Analyse von Restaurant-Daten in 31 europäischen Städten

Bachelorarbeit
Zur Erlangung des akademischen Grades
Bachelor of Science (B.Sc.)
in Volkswirtschaftslehre
an der Wirtschaftswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin

vorgelegt von
Xiaoji Du

Matrikel Nr. 531033

Erstprüferin: Prof. Dr. Sonja Greven

Zweitprüferin: Prof. Dr. Weining Wang

Betreuer: Dr. Sigbert Klinke

Berlin, den 22. Juli 2020

Zusammenfassung

Seit einigen Jahren spielen Bewertungen auf Online-Plattformen im Restaurantsgeschäft eine immer wichtigere Rolle. Immer stärker sind es verfügbare Informationen im Internet, die den Kosum leiten und die Auswahl der Restaurantkunden motivieren. Das Ziel dieser Arbeit ist es, den Einfluss solcher Information zu erkennen und zu messen. Hierzu analysieren wir einen TripAdvisor Datensatz, der Restaurantdaten aus 31 europäischen Städten umfasst. Wir führen die Variablen ein, und visualisieren diese Variablen mit verschiedenen Diagrammerstellungstechniken. Hierfür teilen wir die Analyse in einen Städte-fokussierten und einen Städte-unabhängigen Teil. Für den Städte-fokussierten Teil arbeiten wir heraus, wie sich Städte in wichtigen Variablen unterscheiden. Außerdem präsentieren wir eine Cluster-Analyse, die uns erlaubt, die Ähnlichkeit zwischen Städten als räumliche Nähe darzustellen. In dem Städte-unabhängigen Teil arbeiten wir Korrelationen zwischen Variablen heraus und präsentieren zwei Regressionsmodelle. Eines basierend auf einer multiplen linearen Regression und das andere auf einem Entscheidungsbaum.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

1. Einführung

Das Internet und soziale Medien haben es ermöglicht, sich über Restaurants in Form von Kommentaren, Selbstbeschreibungen und Bewertungen zu informieren. Diese Informationen beeinflussen Kunden in ihrem Verhalten und ihren Konsumentscheidungen. Kunden sind es gewohnt, Onlinemedien zu nutzen um Restaurants nach ihren geschmacklichen und gesundheitlichen Präferenzen und Preisvorstellungen auszuwählen. Zu diesem Zweck sammeln zahlreiche Onlineplattformen und Apps Informationen über Restaurantbesuche und Bewertungen und geben diese an Kunden weiter.

In dieser Arbeit zeigen wir, dass die Beliebtheit von Restaurants mit beobachtbaren Eigenschaften zusammenhängt, die wir der Selbstdarstellung und Kundenbewertung eines Restaurants entnehmen. Im Einzelnen zeigen wir, dass der Bewertungsschnitt eines Restaurants ein ungeeignetes Maß für die Beliebtheit eines Restaurants ist und führen die logarithmierte Anzahl der Besucher als Beliebtheitsmaß ein. Auf dieser Grundlage stellen wir unsere Forschungsfrage: Hängt die Bewertung eines Restaurants mit seiner Selbstdarstellung zusammen? Wenn ja, welche Merkmale treten hierbei am deutlichsten hervor?

Um dies zu zeigen nutzen wir einen öffentlich zugänglichen TripAdvisor Datensatz. TripAdvisor ist eine der größten Onlineplattformen in der Gastronomie. Auf der US-amerikanischen Touristen-Plattform teilen Nutzer ihre Erfahrungen über Unterkünfte, Geschäfte, Sehenswürdigkeiten und Restaurants. Der Tripadvisor Datensatz enthält Restaurantdaten aus 31 europäischen Städten. Er ist aktuell, zuverlässig und umfassend, wobei wir seine Eignung als Stichprobe für eine Grundgesamtheit von Restaurantbesuchen voraussetzen. Ein Beleg dieser Eignung ist außerhalb des Umfangs dieser Arbeit.

Unsere Analyse des Datensatzes teilt sich in drei unabhängige Teile: (i) eine deskriptive Analyse der einzelnen Zufallsvariablen, (ii) einen Städte-orientierten Teil und (iii) einen Restaurant-orientierten Teil. Der deskriptive Teil hilft uns den Datensatz besser zu verstehen.

Im Städte-orientierten Teil fassen wir alle Datensätze zusammen, die in einer Stadt gesammelt wurden. Dann vergleichen wir die so aggregierten Datenpunkte um uns einen Überblick über die Unterschiede und Gemeinsamkeiten der Städte zu verschaffen. Konkret führen wir ein hierarchisches Clustering durch, um die Ähnlichkeiten und Unterschiede zwischen Städten besser zu verstehen. Außerdem führen wir zwei k-means Clusterings durch, in denen die Unterschiede zwischen Städten als Gower Distanz dargestellt wird. Schließlich stellen wir die Verteilung der Selbstbeschreibungslängen der Städte in einem Boxplot dar.

Im Restaurant-orientierten Teil betrachten wir jedes Restaurant bzw. jeden Restaurantbesuch als Beobachtung während wir die Stadt, in der der Datensatz gesammelt wurde, aus der Betrachtung ausklammern. Hierbei interessiert uns wie die einzelnen Zufallsvariablen zusammenhängen. Diese Zusammenhänge erforschen wir mit einer Korrelationsanalyse.

Schlussendlich versuchen wir Regressionsmodelle herzuleiten, mit denen wir die Beliebtheit eines Restaurants mithilfe von anderen beobachtbaren Variablen vorhersagen. Hierfür verwenden wir zunächst ein multiples lineares Regressionsmodell. Anschließend

wenden wir eine CART-Analyse an, die einen Entscheidungsbaum auf Grundlage der Daten entwickelt.

In Abschnitt 2 stellen wir die Methoden vor, die wir zur Visualisierung, Beschreibung, und Modellierung anwenden. Abschnitt 3 stellt den Datensatz vor, welche Variablen ihn charakterisieren und welche Ausprägungen diese Variablen haben. Hiernach folgt die Analyse der Daten. In Abschnitt 4 stellen wir die statistischen Eigenschaften der Variablen des Datensatzes einzeln vor. Es folgt der Städte-fokussierte Teil der Analyse in Abschnitt 5. Abschnitt 6 enthält den Städte-unabhängigen Teil der Analyse. Letztlich geben wir eine Zusammenfassung dieser Arbeit in Abschnitt 7.

2. Methoden

Im Folgenden beschreiben wir kurz die zur Erstellung dieser Arbeit verwendeten Methoden.

2.1. Fehlende Werte: MCAR, MAR, MNAR

In dem Datensatz kommen fehlende Informationen vor. Es wird zwischen drei Arten von fehlenden Werten unterscheiden:

MCAR Missing Completely At Random

Die Werte sind zufällig fehlend. Das Fehlen ist weder von der Variable selbst noch von den unabhängigen Variablen in der Datenbank abhängig. Wenn man sie ignoriert, gibt es weniger Informationen, die Ergebnisse werden jedoch nicht verzerrt.

MAR Missing At Random

Das Fehlen ist abhängig von der eigenen Variable, aber nicht von den anderen Variablen. Bei Ignorieren wird das Ergebnis verzerrt.

MNAR Missing Not At Random

Das Fehlen ist abhängig von anderen Variablen in der Datenbank. Beim Ignorieren wird das Ergebnis verzerrt.[1]

2.2. Spearman'sche Korrelationskoeffizient

Definition Um die Zusammenhänge zwischen den Variablen zu berechnen, verwenden wir den Korrelationskoeffizienten nach Spearman. Der Spearman'sche Korrelationskoeffizient ist definiert in [FKPT11, S.144] durch

$$r_{SP} = \frac{\sum(rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum(rg(x_i) - \overline{rg}_X)^2 \sum(rg(y_i) - \overline{rg}_Y)^2}} \tag{2.1}$$

Wertbereich: $-1 \leq r_{SP} \leq 1$

$r_{SP} > 0$: gleichsinniger monotoner Zusammenhang

$r_{SP} < 0$: gegensinniger monotoner Zusammenhang

$r_{SP} > 0$: kein monotoner Zusammenhang

[1] `https://docs.displayr.com/wiki/Missing_Values`

oder rechentechnisch günstige Version:
bei Daten $(x_i, y_{i)}$, $i = 1, 2, ..., n$, $x_i \neq x_j$, $y_i \neq y_j$ für alle i, j
Rangdifferenzen: $d_i = rg(x_i) - rg(y_i)$

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{2.2}$$

Der Absolutwert des Koeffizienten zeigt die Stärke der Beziehung zwischen den Variablen: Je näher der Wert an 1, desto stärker ist die Beziehung. Ein Koeffizient von 1 zeigt eine perfekte positive lineare Abhängigkeit, während -1 eine perfekte negative lineare Abhängigkeit zeigt. Wenn der Koeffizient 0 beträgt, haben die Variablen keine Beziehung.

Effektsstärke und Signifikanzniveau Wir müssen im Folgenden testen, ob der Korrelationskoeffizient signifikant ist, entspricht ob der p-Wert kleiner als das Signifikanzniveau α ist. Ein α von 0.05 bezeichnet die Wahrscheinlichkeit, dass eine Korrelation rechnerisch bestätigt wird, obwohl tatsächlich keine Korrelation vorhanden wird, 5 % beträgt. Wenn der p-Wert größer als 0.05 ist, entsteht keine aussagekräftige Anzeichen von einer Korrelation zwischen den Variablen.[2]

2.3. Cluster Analyse: Hierarchische Cluster und K-means

Wir haben sowohl hierarchische als auch k-means Klassifikationsverfahren eingesetzt, damit wir die Ähnlichkeit und Distanz zwischen den Städten analysieren können. Um die Ähnlichkeit von n Objekten zu bestimmen, an denen die Merkmale erhoben wurden, werden Ähnlichkeitsmaße verwendet. Die Ähnlichkeitsmaße messen die Ähnlichkeit zwischen dem i-ten und j-ten Objekt. Umso größer das Ähnlichkeitsmaß, desto ähnlicher sind die beiden Objekte (vgl. zu den folgenden Ausführung in [HK17, S.103-105]).

Gower Distanz Unser Datensatz hat numerische und kategoriale Variablen. Mit Gowers Methode können wir nicht nur numerische, sondern auch kategoriale Variablen berücksichtigen. Mathematisch wird die Gower Distanz durch folgende Beziehung in [HK17, S.108] beschrieben

$$d_G(i, j) = \frac{\sum_{k=1}^{p} \delta_{ij}^{(k)} d_{ij}^{(k)}}{\sum_{k=1}^{p} \delta_{ij}^{(k)}} \tag{2.3}$$

für metrische Variablen gilt

$$d_{ij}^{(k)} = \frac{|x_i k - x_j k|}{r_k} \text{ wobei } r_k = \max_i x_{ik} - \min_i x_{ik} \tag{2.4}$$

für kategoriale Variablen gilt

$$d_{ij}^{(k)} = \begin{cases} 1, & \text{wenn } x_{ik} \neq x_{jk} \\ 0, & \text{wenn } x_{ik} = x_{jk} \end{cases} \tag{2.5}$$

[2] `https://support.minitab.com/de-de/minitab/18/help-and-how-to/statistics/basic-statistics/how-to/correlation/interpret-the-results/key-results/`

Mithilfe des Gowers Distanzmaßes ist dann die Ähnlichkeit bzw. die Distanz quantifizierbar, eine Distanzmatrix können wir dadurch für die hierarchische Clusteranalyse bilden.

Complete-Linkage Nach Festlegung der Distanzmaße haben wir agglomerative Verfahren (auch Fusionsalgorithmus genannt) eingesetzt. Unter Verwendung dieses Algorithmuses stellt zu Beginn jede einzelne Stadt einen einzelnen Cluster dar. Die Cluster, die sich am nächsten sind, werden zu einem Cluster zusammengefasst, danach werden die Distanzen bzw. Fusionswerte zwischen den neuen Clustern berechnet. Die Städte werden solange zusammengefasst, bis alle in einem großen Cluster vereint sind.

Als eine Art von Fusionsalgorithmus betrachtet das Complete-Linkage-Verfahren die maximale Distanzen zwischen den Objekten zweier Clustern. Sie wird durch folgende Gleichung beschrieben

$$D_C(C_k, C_l) = \max_{i \in C_k, j \in C_l} d(i,j) \tag{2.6}$$

(vgl. [Wun14, S.44-45])

k-Means Der Mechanismus des k-Means Verfahrens startet mit zufälliger Wahl einer Beobachtung als das erste Cluster-Center, dann wird die minimale Quadratdistanz d_i zu allem Cluster-Centers berechnet. Im Folgenden wird das nächste Cluster-Center gewählt, dass die Wahrscheinlichkeit der Wahl proportional zu d_i ist. Die Berechnung und der Wahlprozess werden wiederholt, bis zum alle k Cluster-Centers festgelegt sind. Basiert auf das Ergebnis wird k-Means Verfahren durchgeführt (vgl. zu den folgenden Ausführungen in [KS16, S.82]).

Bei der K-Means Verfahren muss die Clusteranzahl im Voraus bekannt sein. Meistens muss man bei Festlegung der Clusteranzahl sowohl aus logischer Sicht als auch aus der Forderung der Homogenität heraus, einen Ausgleich finden. Wir können mithilfe des Ellenbogen-Verfahrens die Cluster-Lösung finden.

Ellenbogenkriterium Mit dem Ellenbogenkriterium können wir die Anzahl der Cluster einfach identifizieren. Betrachtet wird die Entwicklung der Heterogenität, die sich durch den Scree-Plot grafisch darstellen lässt. Die x-Achse zeigt die Anzahl der Cluster, die y-Achse zeigt die abgetragene Summe der Fehlerquadrate. Die optimale Anzahl der Cluster finden wir auf dem „Ellenbogen-Knick“ (vgl. [BEW15, S.476]).

2.4. Multiple Lineare Regressionsanalyse

In einem multiplen linearen Modell setzen wir für Y, X_1, ..., X_p die beobachtete Daten in die lineare Funktion ein, so ergibt sich die empirische lineare Beziehung:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i, i = 1, ..., n \tag{2.7}$$

y_i bezeichnet die Zielgröße und die abhängigen Variablen, x_{i1},...,x_{ip} stellen die unabhängigen Variablen dar. β_1, ..., β_p bezeichnen die Regressionskoeffizienten. Sie sind unbekannt und müssen geschätzt werden. Die Anzahl der Beobachtungen lässt sich durch n bezeichnen.

ϵ_i ist die unbeobachtbare Zufallsvariable. Es wird angenommen, dass sie unabhängig und identisch mit $E(\epsilon_i) = 0$ und $Var(\epsilon_i) = \sigma^2$ verteilt sind. Die Annahmen drücken

aus, dass die erwarteten Werte der Fehlerterm Null beträgt und keinen Einfluss auf y_i ausübt, die Fehlervarianz für alle Beobachtungen konstant ist und die Fehler unabhängig und unkorreliert sind.

KQ-Methode

Die kleinste-Quadrate-Methode wird für das Regressionsmodell eingesetzt, um $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_p$ zu schätzen. Ziel der KQ-Methode ist, die geschätzten Regressionskoeffizienten so zu bestimmen, dass

$$\sum_{i=1}^{n} \epsilon_i = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_p x_{ip})^2 \rightarrow \min_{\beta_0, \beta_1, ..., \beta_p} \tag{2.8}$$

ist.

Folgende Voraussetzung müssen in der KQ-Methode erfüllt sein:

1. n muss mindestens so groß sein wie die Zahl der unbekannten Parameter, das heißt $n \leq p + 1$, um den Schätzfehler klein zu halten.

2. Keine Linearkombination unter den erklärenden Variablen darf vorkommen, das heißt es darf für kein j = 0, 1, ..., p

$$X_j = \sum_{k \neq j} a_k X_k + b \tag{2.9}$$

gelten.

Sind die Voraussetzungen erfüllt, erhalten wir die KQ-Schätzer $\hat{\beta}_0$, $\hat{\beta}_1$, ... $\hat{\beta}_p$ dadurch, dass wir die 1. Ableitung nach β_0, β_1,..., β_p gleich null setzen. Von Hand ist die Parameterschätzung in einer multi-linearen Regression nicht durchführbar. Wir fassen die unabhängigen Variablen $x_1 ... x_p$ (als Vektoren angesehen) spaltenweise in eine Matrix X zusammen. Mithilfe dieser Matrix und dem Vektor aller abhängigen Variablen y können wir den Vektor der Parameter schätzen:

$$\beta = (X^T X)^{-1} X^T y \tag{2.10}$$

Das Bestimmtheitsmaß Die Anpassungsqualität des Modells lässt sich durch das empirisches Bestimmtheitsmaß

$$R^2 = \frac{\sum (\hat{Y}_i - \overline{Y_i})^2}{\sum (Y_i - \overline{Y_i})^2} = \frac{SQE}{SQT} = \frac{\text{erklärte Quadratssumme}}{\text{zu erklärende Quadratssumme}} \tag{2.11}$$

bestimmen. Der Wertbereich des Bestimmtheitsmaßes ist $0 \leq R^2 \leq 1$. Je näher R^2 bei 1 liegt, desto besser wird die Zielgröße durch die Regression erklärt.

Die Methoden dieses Abschnittes stammen größtenteils aus [FKPT11, S.494-498]).

2.5. CART-Analyse

CART (Classification and Regression Trees) stellt einen Algorithmus in Form von Baumstrukturen dar, welche als Entscheidungsbäume bezeichnet werden. Entscheidungsbäume sind eine Familie nichtparametrischer Methoden, die wir sowohl für die Regression als

auch zur Klassifikation einsetzen können. Das Ziel von Entscheidungsbäume ist, eine abhängige Variable Y durch unabhängige Variablen X_1,..., X_i vorherzusagen.

Ein Entscheidungsbaum besteht aus Knoten und Ästen und startet mit einem Wurzelknoten. Für jede binäre Entscheidung „ja“ oder „nein“ werden zwei Äste nach dem Top-down-Prinzip gebildet. Die Äste repräsentieren einen Test auf dem Attribut des Knotens und führen zu unterschiedlichen Blättern, welche eine der Klassen repräsentiert. Die Trennkriterien sind, die Entscheidungsknoten so zu wählen, dass möglichst eine homogene Klassenverteilung durchgeführt wird. An der Klassifikation wird die Fähigkeit zur Generalisierung bzw. die korrekte Klassifizierung angefordert (vgl. [Alp10, S.185]).

Komplexitätsparameter CP Eine Anpassung des Entscheidungsbaums müssen wir durchführen. Durch den Komplexitätsparameter CP können wir leicht erkennen, ob der Entscheidungsbaum zu komplex ist.

$$R_{CP} \equiv R(T) + CP * |T| * R(T_1) \tag{2.12}$$

T_1 bezeichnet einen Baum ohne Äste, hier ist der CP-Wert 1, $|T|$ die Anzahl der Ästen und $|R|$ das Risiko. Die Aufteilungen werden solang durchgeführt, bis der Defaultwert von $CP = 0.01$ erreicht wird. Bei jeder weiteren Teilung dieses CP-Wertes kommt es oft vor, dass die Äste überspezialisiert sind, in anderen Worten, das Modell hat ein Überanpassungsproblem (vgl. [TA19, S.12-13]).

Stutzen des Baums Ein derartiges Überanpassungsproblem an die Trainingsdaten lässt sich durch das sogenannte Stutzen (Pruning) vermeiden. Um die Leistung eines Klassifikators abschätzen zu können, teilen wir die Daten in zwei Unterdatensätze: einen Trainingsdatensatz, anhand dessen der Klassifikator aufgebaut bzw. trainiert wird, und einen Testdatensatz, der zur Ermittlung der Klassifikationsleistung verwendet wird. Diese Vorgehen ist für unsere Daten einsetzbar, da unser Datenbestand eine ausreichende Größe aufweist.

3. Datensatz

Für diese Arbeit stellen die von Kaggle Nutzer Damien BENESCHI im 2018 durchgeführte Erhebungen die Datenbasis dar. Der Datensatz umfasst die Informationen über Restaurants auf der bekannten touristischen Webseite TripAdvisor in 31 europäischen Hauptstädten [3]

3.1. Datengrundlagen

Die rohen Daten wurden auf den Listenseiten von TripAdvisor gesammelt, insgesamt wurden Daten von 125.527 Restaurants erhoben. Der Datensatz entspricht keiner Zufallsstichprobe. Bei unserer Analyse wurden folgenden Merkmalen betrachtet:

Name Es handelt sich um die Restaurants, die sich in der TripAdvisor Datenbank registriert haben. Es könnte Restaurants geben, die nicht in der Liste eingetragen sind.

[3] `https://www.kaggle.com/damienbeneschi/krakow-ta-restaurans-data-raw`

City die City-Lage. Die Restaurants befinden sich in 31 europäischen Städten inklusive Amsterdam(NL), Athen (GR), Barcelona (ES), Berlin (DE), Bratislava (SK), Bruxelles (BE), Budapest (HU), Copenhagen (DK), Dublin (IE), Edinburgh (UK), Geneva (CH), Helsinki (FI), Hamburg (DE), Krakow (PL), Lisbon (PT), Ljubljana (SI), London (UK), Luxembourg (LU), Madrid (ES), Lyon (FR), Milan (IT), Munich (DE), Oporto (PT), Oslo (NO), Paris (FR), Prague (CZ), Rome (IT), Stockholm (SE), Vienna (AT), Warszawa (PL) und Zurich (CH).

Cuisine Style die Selbstbeschreibung der Restaurant. Die charakterische Variable enthältet Informationen über die Kochtechnik oder das Ursprungsland des Essens eines Restaurants.

Rating Die Variable informiert uns über die Kundenbewertungen, die auf einer Skala von 1(sehr unzufrieden) bis 5(sehr zufrieden) eingestuft wurden. Die Bewertungsniveaus sind in 11 Stufen mit einem Abstand von 0.5 klassifiziert.[4]

Price Range Die Preise der Restaurants wurden in 3 kategorialen Klassen dargestellt mit $, $$ - $$$ und $$$$. Die Anzahl der Dollar bedeutet hier „Günstiges Essen“, „Mid-Range Essen“ und „Feines Essen“, diese Beschreibungen sind nicht unbedingt von den Preisen abhängig, sondern eher als eine Hinweise für die Kunden. Sie zeigen, ob ein Restaurant über gepflegte Atmosphäre, hohe Kochzutaten-Qualität und guter Service verfügt.[5]

Number of Reviews Anzahl der Bewertungen. Die numerische Variable bezeichnet, wie viele Kunden Rezensionen für die Restaurants abgegeben haben.

Reviews von jedem Restaurant wurden zwei Kundenrezensionen angezeigt.

In der Analyse haben wir folgende zwei neue Variablen erzeugt,

Ltag Die gesamte Zeichenanzahl der Beschreibung von Cuisine-Stil. Diese Variable zeigt die Beschreibungslänge.

Veg eine Dummyvariable. Sie bezeichnet, ob die Schlagwörter „Vegetarian Friendly“ in der Beschreibung des Cuisine-Stil erscheinen.

3.2. Datenaufbereitung

Bevor wir mit der Analyse anfangen, müssen die Daten geprüft, bereinigt und wenn notwendig, transformiert werden. Der Datensatz enthält fehlende Werte, in der Arbeit werden sowohl die Vollständigkeit als auch die Richtigkeit der Datenbank berücksichtigt, um die Realität durch unsere Analyse möglichst genau abzubilden.

[4] https://www.tripadvisor.com/TripAdvisorInsights/w810

[5] https://www.tripadvisor.com/ShowTopic-g1-i12104-k10620761-Restaurant_Price_Classifications-Help_us_make_Tripadvisor_better.html

Datenbereinigung Fehlende Werte, einschließlich der NA-Werte (NA bezeichnet „nicht angegeben“) und leere Zeilen, kommen in unserer Datenbank vor. Zunächst haben wir die leeren Zeilen durch NA-Werte ersetzt, dann haben wir festgestellt, dass es insgesamt 51.302 fehlende Werte in unserer Datenbank gibt. Wie man mit den fehlenden Werten umgeht, ist von der Art abhängig. Die fehlenden Werte in unserer Datenbank bestehen gemischt von MCAR und MAR. Um die Vollständigkeit unseres Datensatzes zu sichern, haben wir in deskriptiver Analyse die fehlende Werte grafisch darstellt und analysiert. Wir untersuchen, ob fehlende Informationen auf die Bewertungsanzahl der Restaurants Einfluss ausüben kann.

Auswahl der abhängigen Variablen Wir haben in der Analyse die Bewertungsanzahl als die Zielgröße eingesetzt. Der Idee liegen folgende Überlegungen zugrunde: Das Bewertungsniveau zeigt uns, was Kunden über ein Restaurant sagen, während die Anzahl der Reviews widerspiegelt, welche Restaurants populär sind, weil die Bewertungsanzahl das tatsächliche Interesse der Kunden zeigt. Durch dieser Auswahlüberlegung verbessern wir die Robustheit unsere Zielgröße gegen Fälschung.

Datentransformation Die Typen der Variablen werden im Weiteren beobachtet. Es gibt in der Datenbank sowohl numerische Variable wie Number of Reviews, als auch kategorische Variable wie Rating und charaktorische Variable wie Price range, City, Cuisine Style und Reviews. Um die Regressionsanalyse und Clusteranalyse durchzuführen, werden entsprechende Variablen wie Price Range von charakteristisch zu kategorisch umgesetzt. Damit die Daten sich normal verteilen und Ausreißer freier werden, haben wir die Bewertungsanzahl durch Logarithmierung transformiert (zur Basis e).

4. Deskriptive Statistik

In diesem Abschnitt analysieren wir den zuvor vorgestellten TripAdvisor Datensatz. Hierzu nutzen wir die Mittel der deskriptiven Statistik um einen groben Überblick über den Datensatz zu gewinnen. In Abschnitt 6 nutzen wir die Erkenntnisse aus der deskriptiven Analyse um in mehreren Korrelationsanalysen festzustellen welche Variablen linear zusammenhängen. Anschließend führen wir eine multiple lineare Regression durch und erstellen einen Entscheidungsbaum. Diese Analysen stellen Zusammenhänge zwischen Variablen her. Hier führen wir zunächst die statistischen Eigenschaften jeder der Variablen für sich ein.

Hierfür stellen wir die Variablen in Tabellen und Graphiken dar. Dadurch können wir die Daten intuitiv erfassen. Wir schätzen Verteilungen, indem wir Häufigkeiten aufsummieren.

4.1. Übersicht

Wir geben zunächst eine Übersicht über die Variablen in dem TripAdvisor Datensatz. Hierfür stellen wir die statistischen Eigenschaften der Variablen in Tabellenform dar.

Tabelle 4.1 zeigt verschiedene Eigenschaften von vier Variablen. Diese Variablen sind für jedes Restaurant die *Anzahl der Bewertungen* (`Num`), das *Bewertungsniveau* (`Rating`), die *Preiskategorie* (`Pcat`), sowie die *Beschreibungslänge* (`Ltags`). Zu jeder dieser Variablen geben wir folgende sieben statistische Eigenschaften: (i) Größe der Stichprobe (n), (ii) Anzahl der Einträge, die für diese Variable undefiniert sind (NAs), (iii) Durchschnittswert (mean), (iv) Standardabweichung (sd), (v) Median (median), (vi) geringste Ausprägung (min), und (vii) höchste Ausprägung (max).

	n	NAs	mean	sd	median	min	max
`Num`	125.527	17.344	125,18	310,83	32,00	2,00	16.478,00
`Rating`	125.527	9.630	3,99	0,68	4,00	0,00	5,00
`Pcat`	125.527	47.855	1,81	0,51	2,00	1,00	3,00
`Ltags`	125.527	31.351	41,55	29,36	34,00	7,00	263,00

Tabelle 4.1: Deskriptive Statistik für die Variablen *Anzahl der Bewertungen* (`Num`), *Bewertungsniveau* (`Rating`), *Preiskategorie* (`Pcat`), und *Beschreibungslänge* (`Ltags`).

Hierin stellen wir fest, dass die *Anzahl der Bewertungen* stark schwankt. Manche Restaurants haben nur 2 Bewertungen. Das meist-bewertete Restaurant hat 16.478 Bewertungen. Ein Median-Restaurant bekommt 32 Bewertungen, während der Durchschnitt sich bei etwa 219 befindet. Der Datensatz lässt eine Unterscheidung zwischen Restaurants ohne Bewertungen und Restaurants mit unbekannter Anzahl der Bewertungen nicht zu. Wir haben deshalb Restaurants ohne Bewertungsangabe als undefiniert interpretiert auch wenn die Vermutung nahe liegt, dass es auch Restaurants ohne Bewertungen geben kann.

Der gesamte Wertebeich des *Bewertungsniveaus* erscheint auch im Datensatz mit sehr schlecht bewerteten Restaurants mit Bewertung 1 bis hin zu sehr gut bewerteten Restaurants mit Bewertungsniveau 5. Das Median-Bewertungsniveau beträgt 4, fast gleich dem durchschnittlichen Bewertungsniveau von $3,99$.

Die Variable *Preiskategorie* hat drei Ausprägungen denen wir die positiven Ganzzahlen 1 bis 3 zuordnen. Das Medianpreisniveau beträgt 2, während das durchschnittliche Preisniveau bei $1,81$ liegt.

Von den 108.178 Restaurants haben etwa ein Viertel ($24,98\%$) über ihren Cuisine-Stil keine Angabe gemacht welche wir als undefiniert interpretieren. In Abschnitt 4.2 zeigen wir, dass die meisten Restaurants nur eine sehr kurze Beschreibung liefern. Nur wenige Restaurants beschreiben sich selbst in vielen Worten.

Im Folgenden betrachten wir zunächst jede Variable einzeln. In Abschnitt 6 ermitteln wir mit Hilfe von Korrelationsanalysen wie die Variablen und die verschiedenen Transformationen dieser Variablen zusammenhängen.

4.2. Restaurant-Selbstbeschreibungen

Das Restaurantsgeschäft ist charakterisiert durch die Wechselwirkungen zwischen Restaurants und Kunden. Diese Wechselwirkungen manifestieren sich in Selbstbeschreibun-

word	freq
vegetarian friendly	32.361
mediterranean	18.428
italian	17.795
vegan	13.009
gluten free	12.120
bar	11.949
french	9.558
asian	9.095
pizza	8.317
spanish	8.220

Tabelle 4.2: Absolute Häufigkeiten der 10 wichtigsten Wörter, die in Restaurant-Selbstbeschreibungen vorkommen

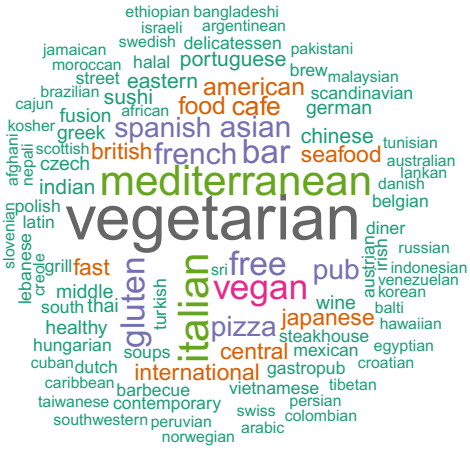


Abbildung 4.1: Wortwolke über die Wörter, die in Restaurant-Selbstbeschreibungen vorkommen

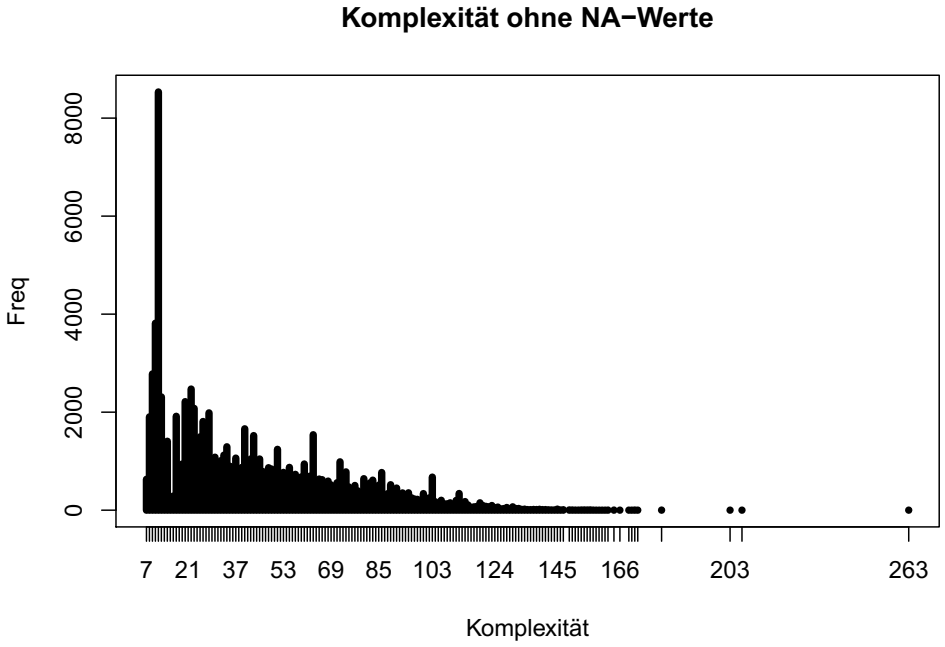


Abbildung 4.2: Histogramm über die Länge der Restaurant-Selbstbeschreibungen

gen, Rezensionen und Bewertungen, die wir beobachten und analysieren können. Wir können uns ad-hoc einen Überblick über die wichtigsten Schlagworte verschaffen indem wir die Wörter, die in den Selbstbeschreibungen eines Restaurants vorkommen, und ihre absolute Häufigkeit erfassen. Tabelle 4.2 gibt einen Überblick über die zehn häufigsten Wörter. Wir haben hierbei Stopwörter herausgefiltert. Um diese Tabelle intuitiv zu verstehen kann man die absoluten Häufigkeiten der Wörter in einer Wortwolke darstellen.

Abbildung 4.1 zeigt die wichtigten Begriffe in der Selbstbeschreibung der Restaurants. „vegetarian friendly“, als häufigstes Schlagwort steht in der Mitte. „mediterranean“ und „italian“ sind die zweit- und dritthäufigsten Begriffe in der Selbstbeschreibung. Interessant ist auch, dass viele Restaurants „vegan“ und „gluten free“ in ihre Beschreibung haben. Insgesamt 32.361 Restaurants beschreiben ihre Kochart mit „vegetarian friendly“, 13.009 Restaurants haben beide Schlagworte „vegan“ und auch „vegetarian friendly“ in ihrer Selbstbeschreibung. Kein Restaurant besitzt „vegan“ aber nicht „vegetarian friendly“ in seiner Selbstbeschreibung. In diesem Datensatz sind somit Restaurants mit veganem Angebot eine echte Untermenge der Restaurants mit vegetarischem Angebot.

Die Wortwolke in Abbildung 4.1 lässt den Rückschluss zu, dass „vegetarian friendly“ das am häufigsten genutzte Schlagwort in Selbstbeschreibungen ist. Basierend auf dieser Beobachtung haben wir die Forschungsfrage festgestellt.

Abbildung 4.2 stellt ein Histogramm über die *Beschreibungslänge*, i.e. die Länge der Selbstbeschreibung eines Restaurants, dar (siehe Anhang A.1). Aus diesem Histogramm geht hervor, dass die meisten Restaurants eine kurze Selbstbeschreibung geben. Bei-

Abbildung 4.3: Wortwolke über die Wörter, die in Bewertungstexten vorkommen

spielsweise haben 8.533 Restaurants eine 11 Zeichen lange Selbstbeschreibung, welches der Mode des Histogramms entspricht. Eher selten geben Restaurants lange Selbstbeschreibungen.

4.3. Bewertungen

Mit der Schlagwortwolke in Abbildung 4.3 fassen wir die häufigsten Wörter in den Bewertungstexten der Restaurantkunden zusammen. Diese Schlagwortwolke ergänzt diejenige, die wir in Abbildung 4.1 für Selbstbeschreibungen der Restaurants erstellt haben.

Wir interpretieren nun die Kundenbewertungstexte auf intuitiver Ebene. „good food“ ist der am häufigsten vorkommende Begriff. „excellent service“ und „great“ werden auch sehr häufig benutzt. Die meisten von den 100 häufigsten Wörtern sind positiv. Wir können dadurch schlussfolgern, dass die meisten englisch-sprachigen Kunden, die auch eine Bewertung schreiben, ihren Restaurantbesuch positiv einschätzen. Zusätzlich stehen uns auch Rezensionsdaten in anderen Sprachen zu Verfügung, z.B. spanisch und chinesisch. In der vorliegenden Arbeit betrachten wir jedoch nur englisch-sprachige Rezensionen.[6] Weiterhin enthält der TripAdvisor Datensatz lediglich zwei Bewertungstexte je Restaurant, unabhängig davon, wie oft ein Restaurant bewertet ist. Unsere Wortwolke würde sich vermutlich anders darstellen, wenn sie auf einem Datensatz erstellt wäre, der alle Bewertungen enthält.

In Abbildung 4.4 betrachten wir die statistischen Eigenschaften der *Anzahl der Bewertungen*. Zunächst können wir an dem Q-Q Plot der untransformierten Variable erkennen,

[6] Das Ergebnis wäre für eine Sentiment-Analyse sehr interessant, wenn die Daten vollständig wären.

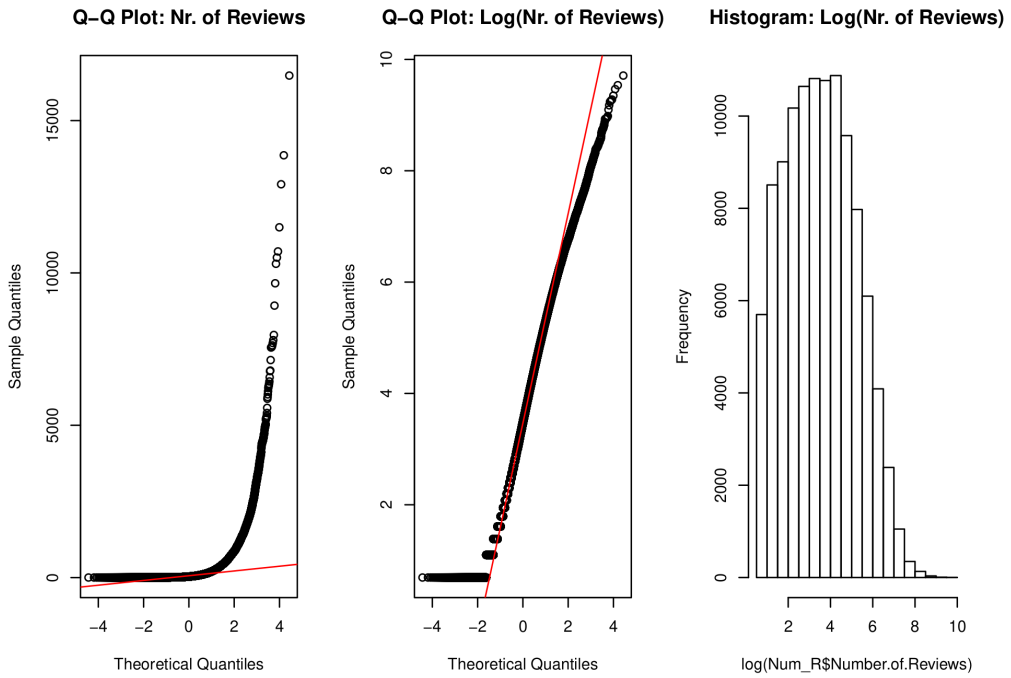


Abbildung 4.4: Q-Q Plots und Histogramm der *Anzahl der Bewertungen*. Links: Q-Q Plot auf der untransformierten Variable. Wir beobachten eine schlechte Übereinstimmung. Mitte: Q-Q Plot über die logarithmierte Variable. Wir beobachten eine zufriedenstellende Übereinstimmung. Rechts: Histogramm der logarithmierten Variable.

dass eine Gauß-Verteilung nicht vorliegt. Jedoch kann eine befriedigende Anpassung an eine Gauß-Verteilung durch Logarithmierung erreicht werden. Zur intuitiven Überprüfung dieser Feststellung zeigen wir auch noch ein Histogramm über der logarithmierten Anzahl der Bewertungen. Das Histogramm zeigt die absoluten Häufigkeiten der verschiedenen Quantile.

5. Städte-Fokussierte Analyse

In diesem Abschnitt betrachten wir den Datensatz mit einem Fokus auf Städte. Die verschiedenen Hauptstädte Europas, die in dem Datensatz auftauchen, unterscheiden sich stark hinsichtlich der Ausprägung des *Preisniveau* und der *Beschreibungslänge*. Daher betrachten wir diese Variablen hier noch einmal für jede Stadt einzeln.

Wir verallgemeinern die Beschreibung unter Anwendung verschiedener Clustering-Verfahren. Wir führen zunächst ein hierarchisches Clustering durch. Im Anschluss konzentrieren wir uns auf durchschnittliches *Bewertungsniveau*, *Anzahl der Bewertungen* und *Preiskategorie*. Mit der Cluster-Analyse können wir die Ähnlichkeiten und Unterschiede durch Distanzen zwischen den Städten ausdrücken.

5.1. Preiskategorien in Verschiedenen Städten

Abbildung 5.1 zeigt die absoluten Häufigkeiten der *Preisniveaus* von Restaurants einzeln für jede erfasste europäische Hauptstadt. Sie erlaubt uns Preisniveaus in verschiedenen Städten zu vergleichen. Je nach Stadt unterscheiden sich die Anzahl der Restaurants. Die meisten erfassten Restaurants befinden sich in London, die wenigsten in Ljubljana. In allen 31 Städten ist eine Mehrzahl der Restaurants auf mittlerem Preisniveau. Es gibt tendentiell mehr günstige Restaurants als teure. Nur in Luxemburg und Zürich ist die Menge der günstigen und teueren Restaurants fast gleich. Der graue Balken zeigt Restaurants, die über ihr Preisniveau keine Angabe machen.

5.2. Beschreibungslänge in Verschiedenen Städten

In Abbildung 5.2 werden die *Längen der Selbstbeschreibung* für die 31 Städte verglichen. Die gruppierte Kastengrafik (Boxplot) zeigt für jede Stadt den Median der Selbstbeschreibungslänge, das mittlere 50% Quantil, das mittlere 95% Quantil, sowie das Minimum und Maximum. Wir sehen, dass die Städte sich klar in Median und Standardabweichung ihrer Beschreibungslänge unterscheiden. Das Restaurant mit der längsten Beschreibung (263 Zeichen) liegt in Amsterdam. Edinburgh hat die höchste Median-Beschreibungslänge. Im Gegensatz hat Lyon die niedrigste Median-Beschreibungslänge. Außer in Oslo, Edinburgh und Dublin kommen in anderen Ländern viele Extremwerte vor, welche wir durch die Punkte erkennen, die außerhalb des 95% Quantils liegen.

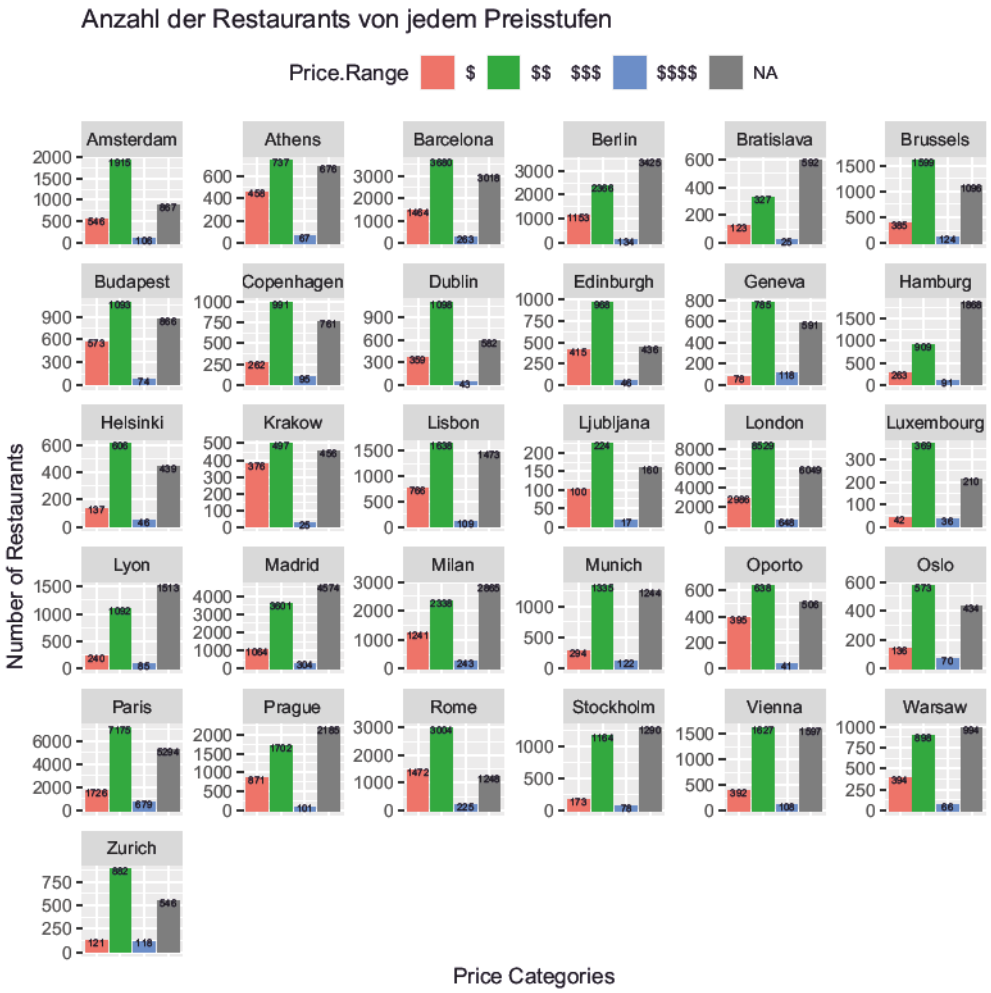


Abbildung 5.1: Häufigkeiten der *Preiskategorien* in verschiedenen Städten

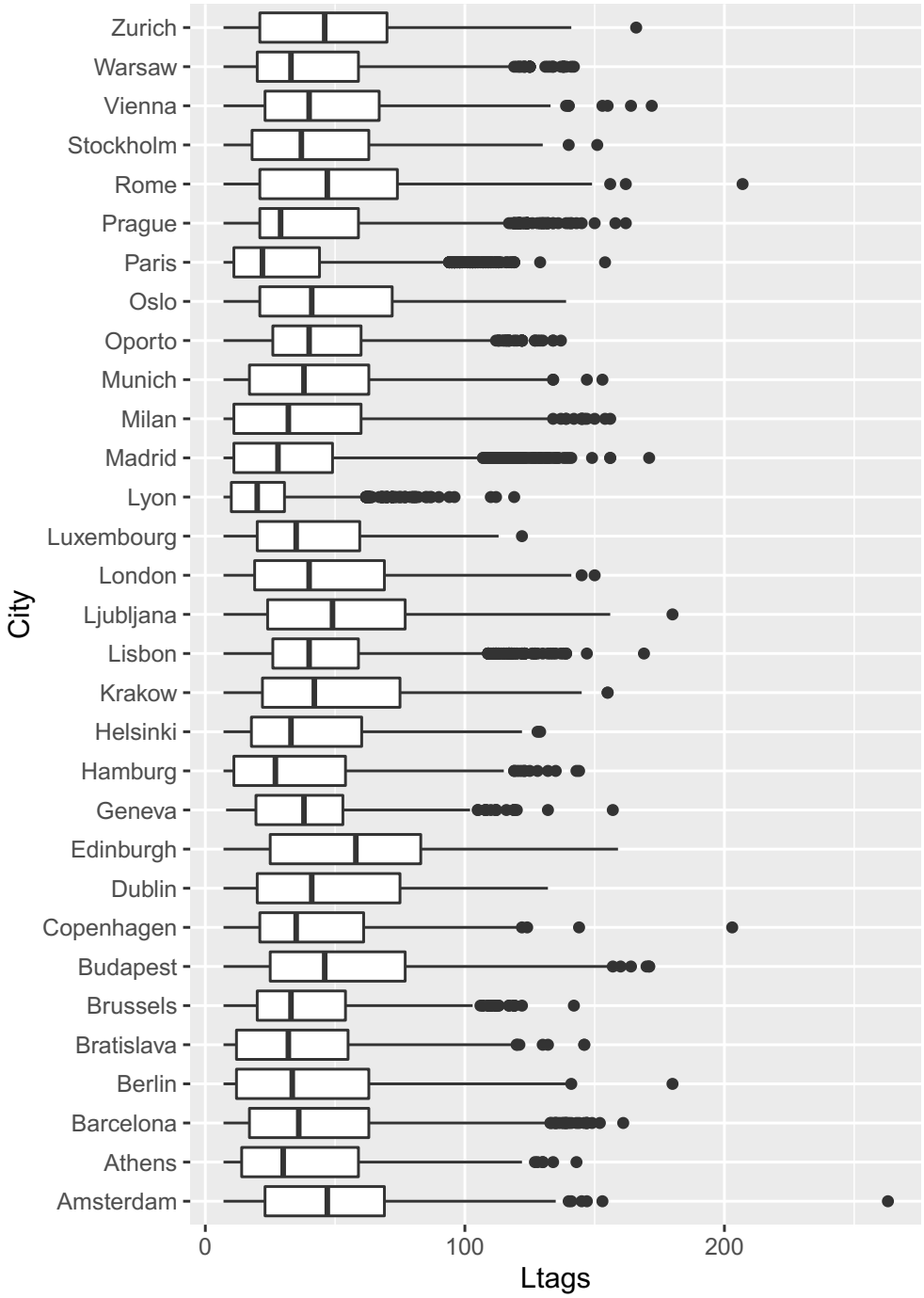


Abbildung 5.2: Boxplot für *Beschreibungslänge* verschiedener Städte

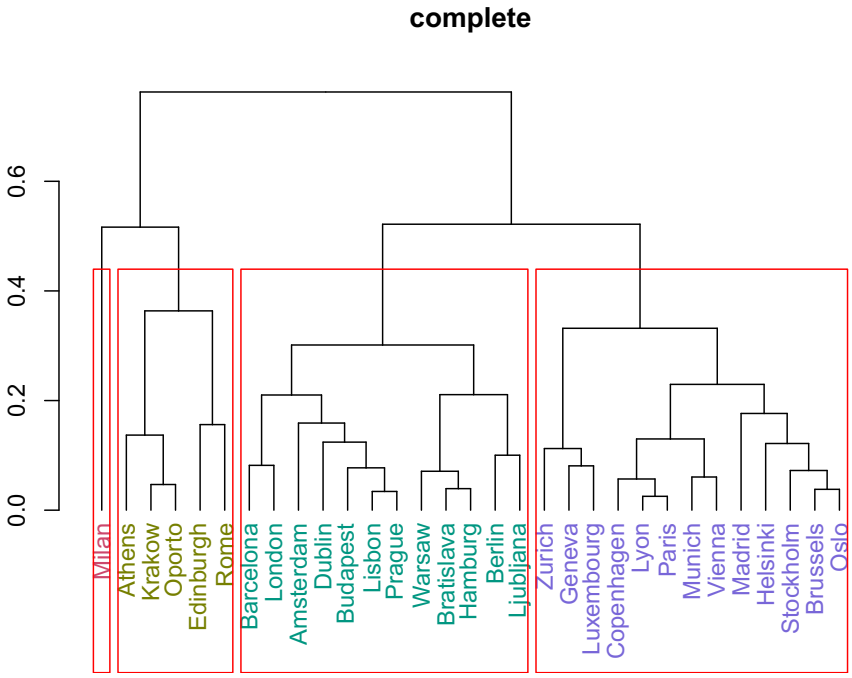


Abbildung 5.3: Hierarchisches Clustering verschiedener Städte

5.3. Hierarchisches Clustering

Als Distanzmaß zwischen Städten nutzen wir die Gower-Distanz. Die Gower-Distanz eignet sich für unseren Datensatz, da es sich sowohl um numerische als auch um kategoriale Variablen handelt. Die Cluster über den maximalen Abstand zwischen den Städten wurden vereint mit dem Complete-Linkage Verfahren.

In Abbildung 5.3 wurden die Gruppenbildung als hierarchische Dendrogramm dargestellt. Der Index auf der y-Achse bezeichnet die Homogenität der Klassen, je kleiner der Index, desto homogener sind die Klassen. Auf der x-Achse befinden sich die 31 Städten. Die vier roten Vierecken umfassen jeweils die Städten eines Clusters. Wir stellen hierbei die Clustern von links nach rechts von 1 bis 4 auf.

Milan ist die einzige Stadt in Cluster 1. Cluster 2 umfasst fünf Städte, die in zwei kleineren Subclustern aufgeteilt wurden. Auf die selbe Art sind 12 Städte in Cluster 3 und 13 Städte in Cluster 4 verbunden. Sehr ähnliche Städte befinden sich in den kleinsten Subclustern.

Nicht nur die hierarchische Struktur, sondern auch die Abstandrelationen zwischen den Städten sind ablesbar in dem Dendrogramm: Paris und Lyon sind hier das Paar der einander ähnlichsten Städte, auf höheren Stufen sind sie mit Copenhagen verbunden. Edinburgh und Rome sind als Paar am höchsten differenziert, weiter gruppiert sind sie mit Athen und mit dem Subcluster aus Krakow und Oporto. Die Gruppierung wird weiterhin fortgesetzt, bis an die oberen Kanten aller vier Vierecke.

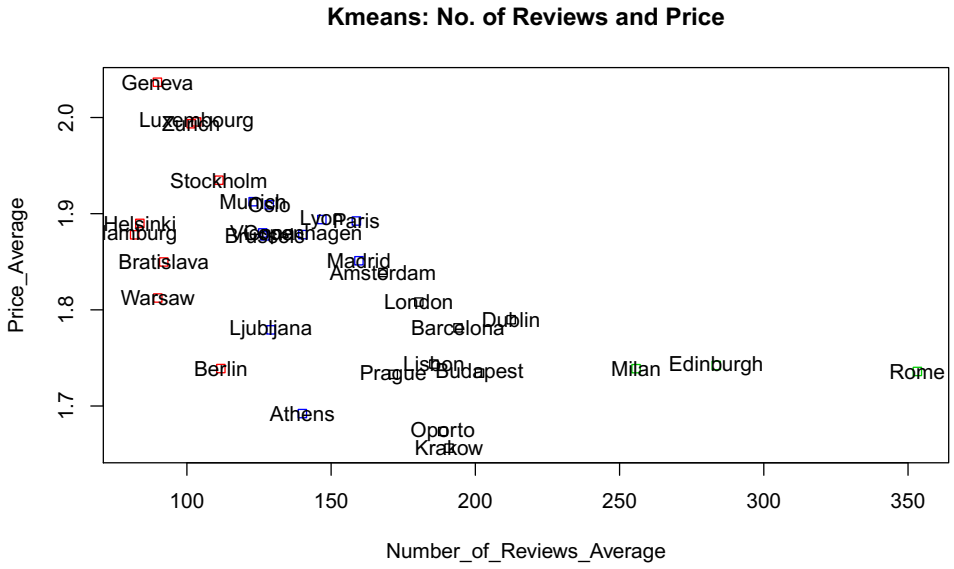


Abbildung 5.4: k-means Clustering für *Anzahl der Bewertungen* und *Preiskategorie*

5.4. K-Means Clustering

Wir nutzen die k-means Methode um Ähnlichkeiten und Unterschiede zwischen den 31 Städten zu finden und zu visualisieren. Das k-means Clustering erlaubt uns Städte mit ähnlichen durchschnittlichen *Preis*, *Bewertungsniveaus* oder *logarithmierter Bewertungsanzahl* zu Gruppen zusammenfassen. In jeder Gruppe sind die Städte sich untereinander eher ähnlich. Im Umkehrschluss sind Städte in unterschiedlichen Gruppen eher verschieden.

Abbildung 5.4 zeigt das Clustering auf den Variablen durchschnittliche *Anzahl der Bewertungen* und *Preis*. Wir sehen, dass die Anzahl der Bewertungen in Städten mit hohen Preisen, wie Geneva, Zürich und Luxemburg, eher gering ist. Städte mit einer hohen Bewertungsanzahl, wie Rom, Edinburgh oder Mailand, haben niedrige durchschnittliche Preise. Niedrigere Preise führen aber nicht unbedingt zu einer höheren Anzahl der Bewertungen. Beispielsweise hat Berlin sowohl ein geringes Preisniveau als auch eine geringe Anzahl der Bewertungen.

Abbildung 5.5 zeigt das Clustering auf den Variablen durchschnittliches *Bewertungsniveau* und *Preis*. Das durchschnittliche Bewertungsniveau der meisten Städte liegt zwischen 3.9 und 4.2. Die niedrigste durchschnittliche Bewertung beobachten wir in Mailand, obwohl in Mailand der Durchschnittspreis niedrig ist.[7]

In Verbindung mit der Beobachtung in den Abbildung 5.4 und Abbildung 5.5 sehen wir, dass Stockholm, Oslo, Brüssel, Helsinki und Madrid als teure Städte relativ wenige

[7] Zu bemerken ist, dass die Preiskategorie nicht objektiv erhoben wird. Somit kann eine niedrige Preiskategorie in beispielsweise Mailand etwas völlig anderes bedeuten als in Ljubjana.

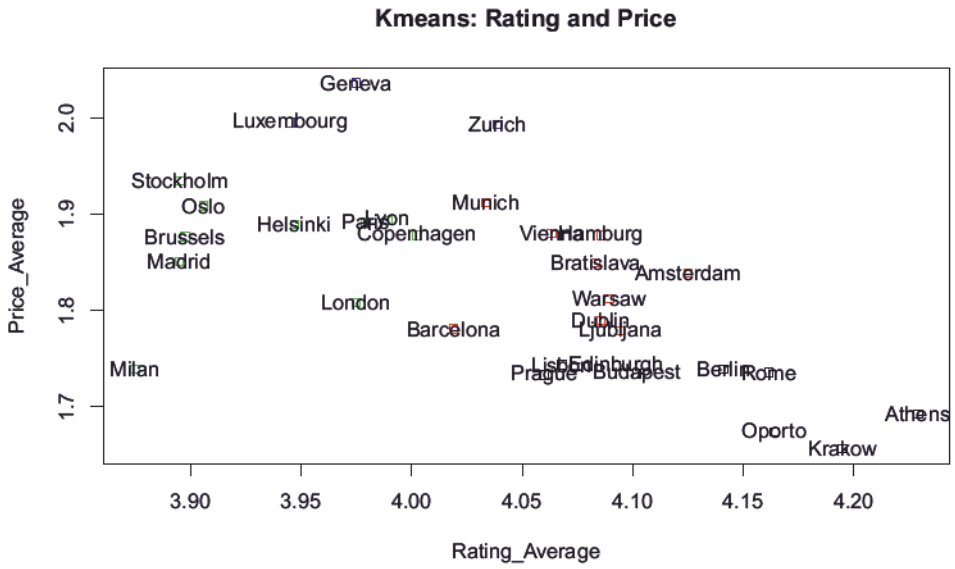


Abbildung 5.5: k-means Clustering für *Bewertungsniveau* und *Preiskategorie*

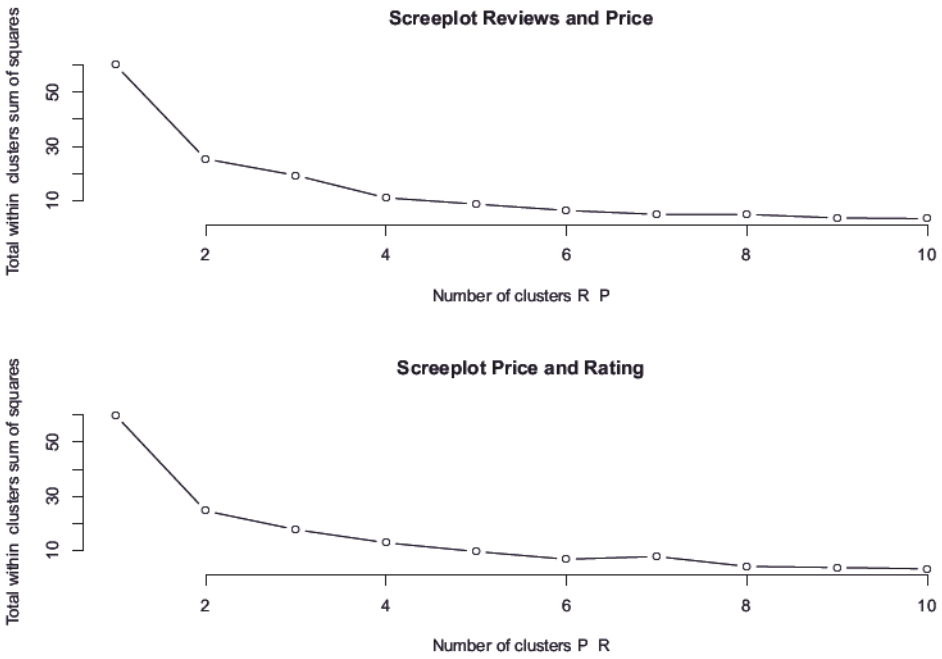


Abbildung 5.6: Summe der quadratischen Fehler abhängig von der Clusteranzahl

Bewertungsanzahlen und niedrige Bewertungen haben. Zu der gut bewerteten Gruppe gehören Berlin, Rom, Oportom, Krakau und Athen, alle obige Städte sind auf niedrigen Preisniveaus. Rom ist die einzige Stadt, die sowohl viele Bewertungsanzahl hat, als auch sehr gut bewertet wurde.

Mit Hilfe der Ellenbogen Methode bestimmen wir zunächst die optimale Anzahl der Cluster. In Abbildung 5.6 sehen wir, dass der "Ellbogen-Knick" in beiden Grafiken, auf 4 liegen. Wir können diese Anzahl als Cluster-Lösung verwenden.

6. Städte-Unabhängige Analyse

Bei der Korrelationsanalyse verwenden wir den Spearman'schen Korrelationskoeffizienten, um den Zusammenhang zwischen den Ausprägungen "vegan-friendly" und "not vegan-friendly" mit (i) der Bewertung und (ii) der Anzahl der Bewertungen zu messen.

Weiterhin bauen wir drei lineare Regressionsmodelle auf, um die lineare Abhängigkeit zwischen folgenden Variablen zu untersuchen: (i) „Anzahl der Bewertungen" in Abhängigkeit von der „Bewertungsniveau", was Rückschlüsse darüber zulässt ob gut bewertete Restaurants auch häufiger besucht werden, (ii) „Bewertungsniveau" in Abhängigkeit vom „Preiskategorie", was Rückschlüsse darüber zulässt wie Preispräferenz und Erwartungshaltung der Kunden deren Bewertungen beeinflussen, und (iii) „Anzahl der Bewertungen" in Abhängigkeit von der „Selbstbeschreibungslänge" was Rückschlüsse darüber zulässt wie stark die Besucherzahlen eines Restaurants von der Elaboriertheit ihrer Selbstbeschreibung abhängt.

Ausgehend von dem Ergebnis der linearen Regression, untersuchen wir die Vorhersagekraft eines CART-Modells in das alle vorher untersuchten Variablen einbezogen werden. Abschließend prüfen wir die Anpassungsgüte des Modells.

6.1. Korrelation zwischen Bewertungsniveau und Preiskategorie

Den Spearman'sche Korrelationskoeffizient verwenden wir zunächst, um den tatsächlichen Zusammenhang zu analysieren. Laut dem Ergebnis beträgt der Koeffizient bei Anzahl Bewertungen und Bewertungsniveau -0.051, es entsteht daher fast keine Korrelation zwischen den beiden Variablen. Bei Bewertungsanzahl und Preisniveau wird eine schwache Korrelation aufgewiesen, der Koeffizient beträgt 0.2753.

Aus der Betrachtung der linken Grafik in Abbildung 6.1 werden die Restaurants auf niedrigeren Bewertungsstufen 1 bis 3 wesentlich weniger bewertet. Die Restaurants mit viele Bewertungen verteilen sich zwischen 3.5 und 4.5, Restaurants auf Niveau 4 bekommen die meisten Bewertungen. Restaurants auf Niveau 5 werden wieder weniger bewertet, trotz des höchsten Bewertungsniveaus. Der Zusammenhang zwischen Bewertungsniveau und Preisniveau wird in der rechten Grafik klar gezeigt. Restaurants auf dem Preisniveau mid-Range haben die meisten Bewertungen. Die Verteilung der Bewertungen auf jedem Preisniveau sind ähnlich verteilt wie in der linken Grafik. Die Restaurants zwischen Bewertungsniveau 3.5 und 4.5 werden am meisten bewertet.

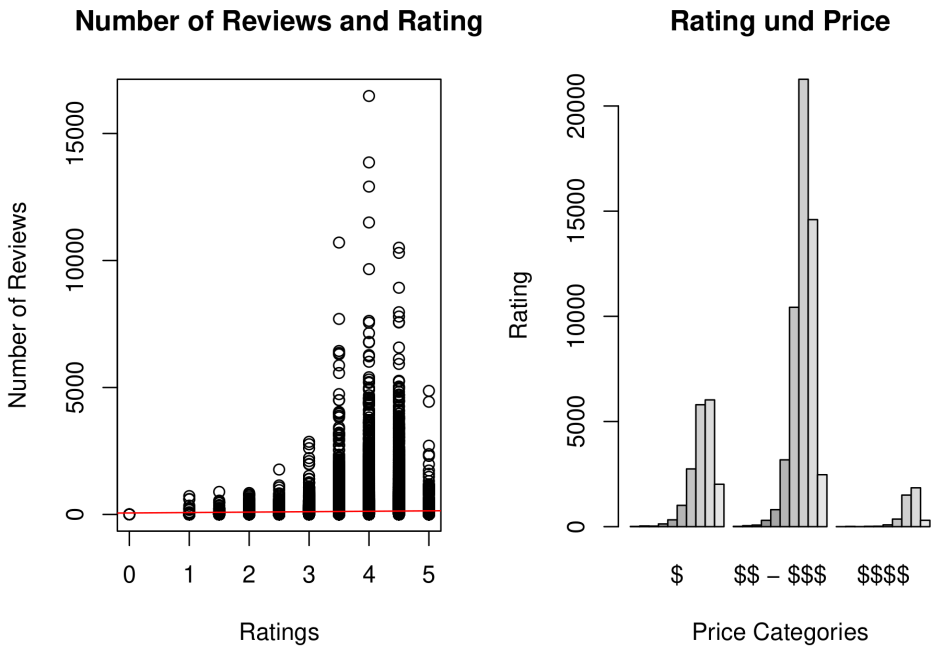


Abbildung 6.1: Korrelation: *Anzahl der Bewertungen* vs *Bewertungsniveau* (links); *Anzahl der Bewertungen* vs *Preiskategorie* (rechts)

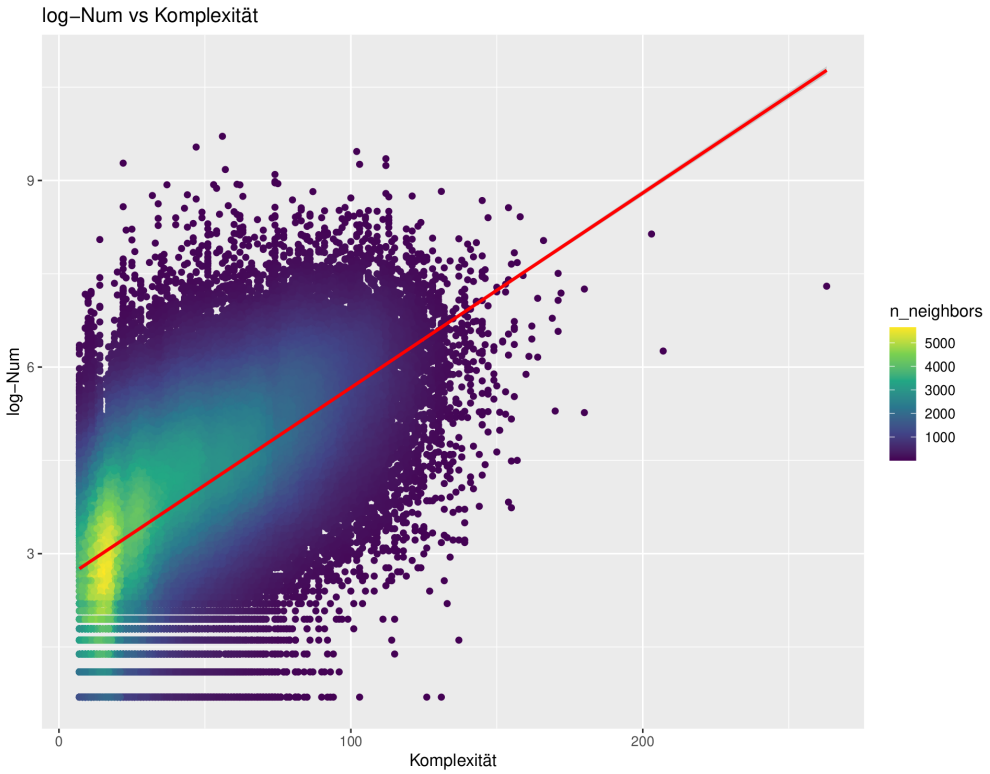


Abbildung 6.2: Korrelation: *log-Anzahl der Bewertungen* vs *Beschreibungslänge*

6.2. Korrelation zwischen Anzahl der Bewertungen und Beschreibungslänge

Der Koeffizient zwischen Beschreibungslänge und logarithmierter Anzahl der Bewertungen beträgt 0.6002. Die steigende rote Linie in Abbildung 6.2 veranschaulicht die stark positive Korrelation. Zwischen Beschreibungslänge und Bewertungsniveau kommt ein Koeffizient von 0.166 vor, welche eine schwache positive Korrelation aufweist.

Für die Signifikanz der obigen Korrelationsbeziehungen liegen aussagekräftigen Anzeichen vor, denn alle p-Werte sind fast 0 (p-value $<$ 2.2e-16).

6.3. Korrelation zwischen Anzahl der Bewertungen und Vegetarierfreundlichkeit

Der Mosaicplot in Abbildung 6.3 stellt den Mehrwege-Zusammenhang zwischen den kategorialen Variablen Preisniveau, Vegetarier-freundlich, und in 10 Kategorien geteilt Anzahl der Bewertungen.

Aus der Betrachtung der Grafik kommen auf jedem Preisniveau die nicht-Vegetarierfreundlichen Restaurants bei Bewertungsanzahl Stufe 1 bis 5 überwältigend häufiger als die Vegetarier-freundliche Restaurants vor. Bei Bewertungsanzahl Stufe 6 bis 10 umgekehrt.

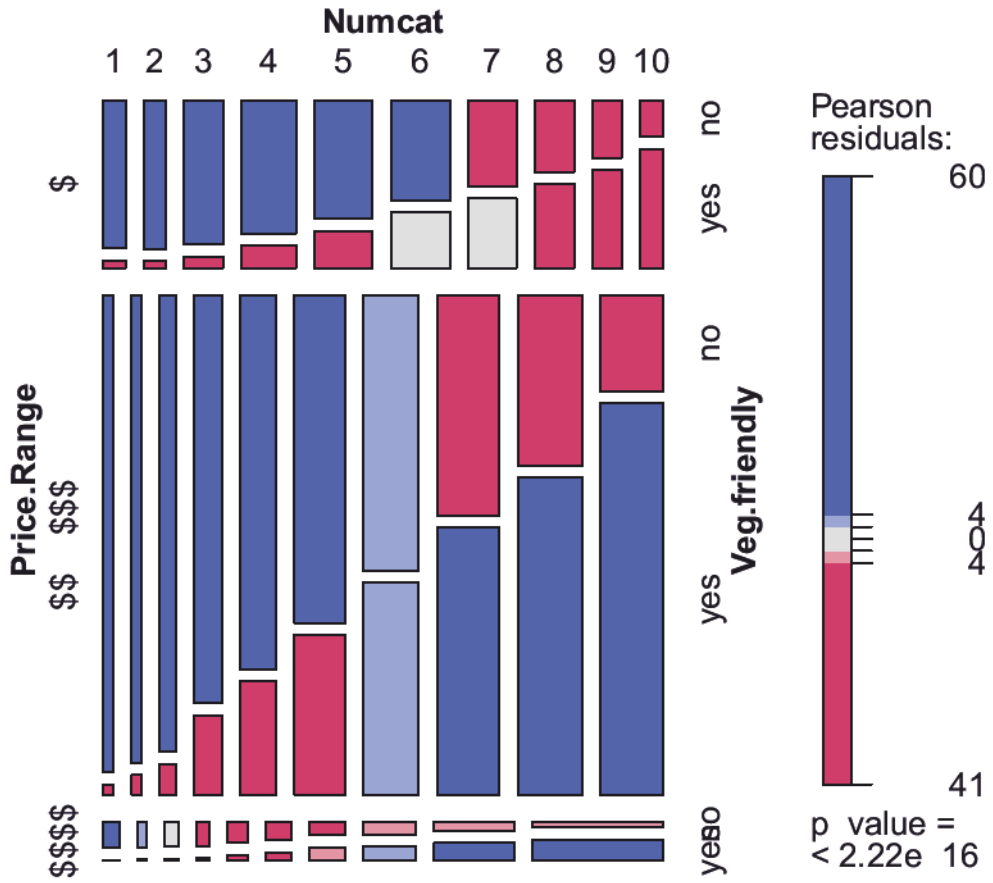
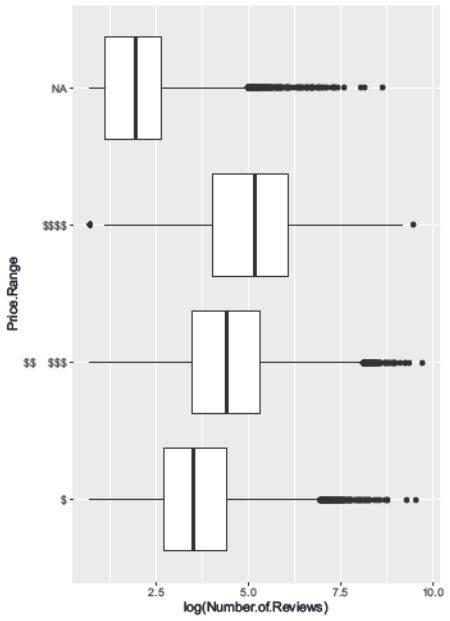
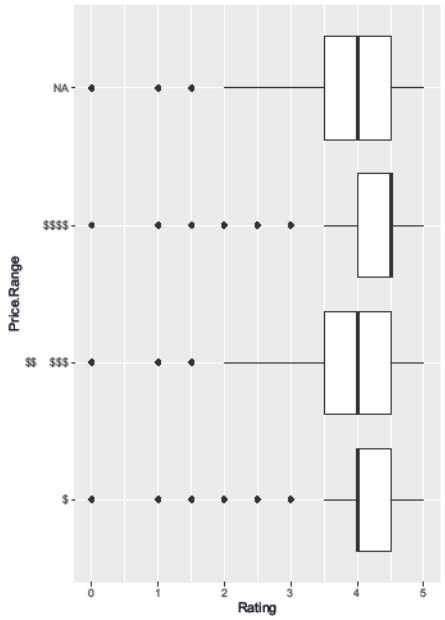


Abbildung 6.3: Mosaicplot über Korrelation zwischen *Preiskategorie*, *Anzahl der Bewertungen* (in 10 Kategorien) und *Vegetarier freundlich*



(a) Boxplot von log-Anzahl der Bewertung und Preisniveaus



(b) Boxplot von Rating und Preisniveaus

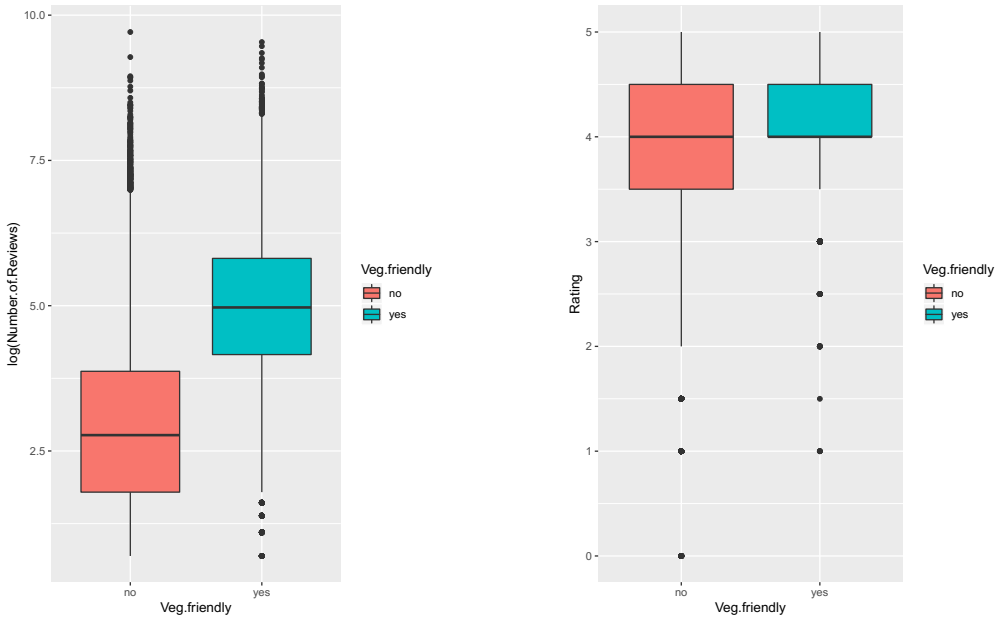
Abbildung 6.4: Zusammenhang zwischen *Preiskategorie* und *log-Anzahl der Bewertungen* sowie *Bewertungsniveau*

Die Stufen der Residuale lassen sich von den Farben her unterscheiden. Gezeigt werden sie auf der rechte Seite des Mosaicplots. Blau bedeutet, dass mehr Beobachtungen als erwartet wurden unter dem Nullmodell (Unabhängigkeit), rot bedeutet das Gegenteil. Hier beträgt die blaue Pearson Residuale 60 und die rote -41. In einem guten Modell sollte die Pearson Residuals klein sein, das heißt wenige oder keine Farbe. Hier ist das nicht der Fall. Die P-Werte von fast Null, kleiner als das Signifikanzniveau von 0.05, weist ein signifikantes Chi-Quadrat-Test Ergebnis auf. [ZMH07]

Korrelation zwischen Preiskategorie und Bewertungsniveau

Wir möchten wissen, ob die *Preiskategorie* Rückschlüsse auf *Bewertungsniveau* oder *Anzahl der Bewertungen* zulassen. Abbildung 6.4a zeigt, dass sich die logarithmierte Anzahl der Bewertungen je nach Preiskategorie unterscheiden. Je teurer ein Restaurant ist, desto höher ist auch der Median seiner Besucherzahl. Der Boxplot für die höchste Preiskategorie weist offenbar einen breiteren Interquartilsabstand und weniger Ausreißer im Vergleich mit den anderen Preiskategorien auf. Der graue Kasten bezeichnet fehlende Werte. Kunden bewerten offenbar Restaurants ohne Preiskategorie eher wenig.

Abbildung 6.4b stellt den Unterschied zwischen der Zusammenhängen von *Preiskategorie* und *Bewertungsniveau* dar. Ein Zusammenhang zwischen beiden Variablen ist augenscheinlich nicht gegegeben. Der Interquartilsabstand vom Preisniveau „günstiges



(a) Boxplot von log-Anzahl der Bewertungen und Vegetarier-freundlich

(b) Boxplot von Bewertung und Vegetarier-freundlich

Abbildung 6.5: Zusammenhang zwischen *Bewertung* und *Vegetarier-freundlich*

Essen“ ist auf der selben Höhe wie der vom Preisniveau „feines Essen“ zwischen Bewertungsniveau 4 und 4.5, jedoch ist der Median von günstigen Restaurants auf dem Boden des unteren Quartil bei Bewertungsniveau 4, anders als der Median von feinen Restaurants, der auf dem Himmel des oberen Quartil bei Bewertungsniveau 4.5 liegt. Sowohl der Interquartilsabstand als auch die Mediane von „Mid-Range“ Restaurants, ist auf der gleiche Höhe wie bei Restaurants ohne Information bei Bewertungsniveau 4.

Die extremen Werte der logarithmierten Anzahl der Bewertungen sind fast alle über dem oberen Whisker, das bedeutet, extreme höhere log-Anzahl der Bewertungen auf jedem Preisniveau kommen vor. Ganz im Gegenteil weist jede Preisstufe extrem niedrige Bewertungsniveaus auf.

6.4. Korrelation mit Vegetarierfreundlichkeit

Um den Unterschied zwischen der Bewertung von Restaurants mit und ohne „Vegetarian friendly“ darzustellen, betrachten wir die Kastengrafiken in Abbildung 4.6. Die linke Grafik zeigt, dass die log-Bewertungsanzahl von Restaurants mit „Vegetarier-freundlich“ wesentlich höher sind, als die ohne, sowohl der Interquartilsabstand als auch die Mediane. Der Unterschied zwischen den Bewertungsniveaus beider Gruppen wird in der rechte Grafik veranschaulicht: der Median der Vegetarier-freundlichen Restaurants liegt auf dem Boden des Kastens, während sie eine um die Hälfte kürzeren Interquartilsabstand als die nicht-Vegetarier-freundliche haben. Der Median von nicht-Vegetarier-freundliche

Restaurants liegt in der Mitte des Kastens und auf Niveau 4, gleich wie bei den anderen Gruppen.

Extreme Werte kommt in beiden Grafiken vor. Betrachtet man die linke Grafik, sind alle extremen Werte bei Restaurants ohne „Vegetarian friendly“ über dem oberen Whisker. Bei Restaurants mit „Vegetarian friendly“ kommen die extremen Werte sowohl über dem oberen als auch unter dem unteren Whisker vor. In der rechte Grafik beobachten wir nur wenige extremen Werte unter beiden unteren Whiskern.

6.5. Vorhersage durch Multiple Lineare Regression

Im Folgenden stellen wir ein multiples, lineares Regressionsmodell über die Zufallsvariablen unseres Datensatzes auf die wir betrachten und die sich numerisch darstellen lassen. Wir versuchen also, die Auswirkungen mehrerer Variablen auf eine Zielvariable als lineare Abhängigkeit darstellen und somit als Hyperebene in einem n-dimensionalen Raum, wobei n die Anzahl der betrachteten Variablen einschließlich der Zielvariable ist.

Als abhängige Variable wählen wir die logarithmierte *Anzahl der Bewertungen* (`Num`). Dies folgt aus unserer Einschätzung, dass die Anzahl der Bewertungen ein aussagekräftiger Maßstab für Beliebtheit eines Restaurants ist. Dies ist in unserer Vermutung begründet, dass die Anzahl der Bewertungen proportional ist zu der Anzahl der Besucher eines Restaurants. Wir logarithmieren diese Variable aufgrund der Beobachtung, dass dadurch die abhängige Variable eher normal verteilt ist (siehe Abschnitt 6 und Abbildung 4.4).

Als Einflussvariablen wählen wir die kategorielle Variable *Bewertungsniveau* (`Rating`), die diskrete Variable *Beschreibungslänge* (`Ltags`), die kategorielle Variable *Preiskategorie* (`Pcat`) und die Boolsche Variable *Vegetarier-freundlich* (`Veg`), die wir als Dummyvariable in das lineare Regressionsmodell einbringen. Die Eigenschaften des Modells stellen wir in Tabelle 6.1 dar. Das zugehörige R Listing ist in Anhang A.2 dokumentiert.

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	3.9179	0.0348	112.52	0.0000
`Rating`	-0.4218	0.0078	-54.31	0.0000
`Pcat`	0.4555	0.0084	54.40	0.0000
`Ltags`	0.0194	0.0002	92.65	0.0000
`Veg`	0.5547	0.0124	44.77	0.0000

Tabelle 6.1: Eigenschaften des multiplen linearen Regressionsmodells für die logarithmierte *Anzahl der Bewertungen* (log(`Num`)) abhängig von *Bewertung* (`Rating`), *Preiskategorie* (`Pcat`), Beschreibungslänge (`Ltags`) und *Vegetarier-freundlich* (`Veg`). Die linke Spalte (Estimate) zeigt die Parameterschätzung für die gegebene Einflussvariable.

Aus den Parameterschätzungen in Tabelle 6.1 können wir das folgende lineare Regressionsgleichung aufstellen:

$$\log(\texttt{Num}) = 3.9179 - 0.4218\ \texttt{Rating} + 0.4555\ \texttt{Pcat} + 0.0194\ \texttt{Ltags} + 0.5547\ \texttt{Veg} + \epsilon \quad (6.1)$$

Im Folgenden erklären wir nocheinmal die Variablen, die in der multiplen linearen Regression erscheinen:

$\log(\texttt{Num})$ logarithmierte Anzahl der Bewertungen als nicht-negative reelle Zahl im Intervall $[0, \infty)$.

`Rating` Bewertungsniveau als positive Zahl im Intervall $[1, 5]$ diskret fortschreitend in Schritten von 0.5, was neun Ausprägungen zulässt.

`Pcat` Preiskategorie als positive Ganzzahl im Intervall $[1, 3]$

`Ltags` Beschreibungslänge als nicht-negative Ganzzahl im Intervall $[0, \infty)$.

`Veg` mit oder ohne „Vegetarian Friendly“ in der Beschreibung als Ganzzahl im Intervall [0,1]

ϵ Zusammenfassung aller unbeobachtbaren Zufallsprozesse, die unabhängig und identisch verteilt sind mit $\mathbb{E}(\epsilon) = 0$ und $\text{Var}(\epsilon) = \sigma^2 | \sigma^2 \in \mathbb{R}^+$

Interpretation

Alle P-Werte für das Modell liegen bei 0.0000 und sind somit kleiner als ein Signifikanzniveau $\alpha = 0.01$. Dadurch erkennen wir, dass die multi-lineare Regression zwischen der abhängigen und unabhängigen Variablen signifikant ist. Somit entscheiden wir uns, alle unabhängigen Variablen in dem Modell zu belassen. Das empirische Bestimmtheitsmaß $R^2 = 0.3629$ bedeutet, dass 36.29% der Variation in der logarithmierte Bewertungsanzahl durch das Modell erklärt werden können. Der y-Achsenabschnitt (Intercept) beträgt 3.9179. Diesen zu interpretieren hätte nur in einem Szenario Sinn, in dem es eine Preiskategorie 0 und ein Rating von 0 gäbe. Wir haben uns jedoch dafür entschieden, beide Variablen von 1 an zu belegen. Somit unterbleibt eine Interpretation des y-Achsenabschnittes.

Die geschätzten Regressionskoeffizienten stellen dar, wie sich die abhänge Variable bei einer Einheit Veränderung der unabhängigen Variable ändert. Beispielsweise wenn man die Bewertung um eine Stufe erhöht, sinkt die logarithmierte Bewertungsanzahl um 0.4218. Erhöht man die Preiskategorie um eine Stufe, so erhöht sich die logarithmierte Bewertungsanzahl um 0.4555. Jedes Wort, das der Selbstbeschreibung hinzugefügt wird führt zu einem Anstieg der logarithmierte Bewertungszahl um 0.01936. Wenn ein Restaurant in seiner Selbstbeschreibung die Formel „Vegetarian Friendly“ nutzt, steigt die logarithmierte Bewertungsanzahl um 0.5547. Das impliziert, dass sich bis auf das Bewertungsniveau alle unabhängigen Variablen positiv auf die Zielgröße auswirken. Jedoch

> [...] hängt die Signifikanz mit der Fallzahl zusammen. Bei hohen Fallzahlen können auch kleine Unterschiede (bzw. schwache Zusammenhänge) signifikant werden – auch wenn diese Unterschiede inhaltlich kaum relevant sind.[8]

[8] `https://statistik-dresden.de/archives/857/`

In Abschnitt 6 haben wir bereits festgestellt, dass Bewertungsniveau und logarithmierte Anzahl der Bewertungen fast unkorrelliert sind (siehe Abbildung 6.1). Daraus folgt auch dass ihr Platz in einem linearen Regressionsmodell fraglich ist. Es ist möglich, dass es andere Zufallsvariablen in oder außerhalb des von uns betrachteten Datensatzes gibt, die die Vorhersage der Zielgröße erheblich verbessern. Eine erschöpfende Aufzählung dieser möglichen Variablen und ihrer Transformationen liegt jedoch außerhalb des Umfangs dieser Forschungsarbeit.

6.6. CART: Klassifikation und Regression

Bisher haben wir in unserem Datensatz nach linearen Zusammenhängen gesucht. Die Korrelationsanalyse, die Clusteranalyse, sowie auch die lineare Regressionsanalyse sind ausschließlich in Szenarien sinnvoll in denen entweder lineare Zusammenhänge bestehen oder solche durch einfache Transformationen hergestellt werden können. Mit der CART-Analyse steht uns ein Werkzeug zur Verfügung, das so allgemein ist, dass es auch nicht-lineare Zusammenhänge innerhalb der erfassten Variablen entdeckt.

Hier führen wir eine Klassifikationsregression durch, welche zum Ergebnis einen Entscheidungsbaum über den TripAdvisor Datensatz hat. Die Größe der Stichprobe erlaubt uns eine n-fache Kreuzvalidierung mit hinreichend großen Partitionen durchzuführen. Abbildung 6.8 zeigt den resultierenden Entscheidungsbaum. Der Baum ist das Ergebnis eines Stutzungsprozesses (Pruning) der durch den Generalisierungsfehler aus der n-fachen Kreuzvalidierung geleitet ist.

Komplexitätsparameter und R^2 Gütemaß

Die Güte eines Entscheidungsbaums kann man durch das R^2 Gütemaß ausdrücken. Wir suchen zunächst einen Entscheidungsbaum mit möglichst hohem R^2 Wert. Man kann jedoch feststellen, dass sich der R^2 Wert immer weiter verbessern lässt, indem man den Entscheidungsbaum um Entscheidungen erweitert. Dadurch kommt es ab einem bestimmten Punkt zu einer Überanpassung. Diese Überanpassung ist unerwünscht.

Einen Entscheidungsbaum um weitere Entscheidungen zu erweitern erhöht seine Komplexität. Wir können die Komplexität eines Entscheidungsbaums durch die Anzahl seiner Endknoten ausdrücken. Weiterhin können wir unseren Datensatz einteilen in einen Trainingsdatensatz und einen Testdatensatz. Diese Einteilung erlaubt es, den Entscheidungsbaum auf dem Trainingsdatensatz zu erstellen und seine Güte auf dem Trainingsdatensatz zu erfassen. Von hier ab nennen wir das Gütemaß auf dem Trainingsdatensatz den Trainings-R^2. Weiterhin haben wir nun die Möglichkeit die Güte auch für den Testdatensatz zu erfassen. Da dieser Datensatz für das Training nicht benutzt wurde ist er neu und wir können beobachten wie gut der Entscheidungsbaum generalisiert. Von hier ab nennen wir das Gütemaß auf dem Testdatensatz den Generalisierungs-R^2. Die n-fache Kreuzvalidierung partitioniert die Stichprobe in n Partitionen, die je einmal als Testdatensatz verwendet werden. Dies erlaubt es einen Schnitt über den Trainings-R^2 und den Generalisierungs-R^2 zu ziehen. Wenn wir die n-fache Kreuzvalidierung für verschiedene Baumkomplexitäten durchspielen stellen wir fest, dass bis zu einem bestimmten

Punkt der Generalisierungs-R^2 genau wie der Trainings-R^2 zunimmt. Für höhere Komplexitäten nimmt der Generalisierierungs-R^2 ab während der Trainings-R^2 weiter steigt. Dieser Punkt, in dem der Generalisierungs-R^2 sein Maximum hat, bestimmt die optimale Baumkomplexität.

Wir führen den CART Algorithmus inklusive einer n-fachen Kreuzvalidierung mit dem R^2 Wert als Gütemaß durch rsq.rpart und erhalten die Tabelle 4.3 sowie zwei Bilder in Abbildung 4.15. In Tabelle 4.3 stellen wir die Komplexitätsparameter CP von der Klassifikationsregression dar. *nsplit* bezeichnet die Anzahl der Teilungen. *rel error* bezeichnet das empirische Ergebnis von $1-R^2$ und *xerror* der Generalisierungsfehler (cross-validated error rate). *xstd* ist der Standardfehler des Generalisierungsfehlers (standard Error). CP = 0.01 ist der Defaultwert. Bis zu diesem Wert beendet sich die Entwicklung des Baumes, ein Überanpassungsproblem taucht aber oft auf. Um den optimalen Komplexitätsparameter zu finden, folgen wir der in [Sch13, S176-177] beschriebenen 1-SE Regel: die Minimalstelle liegt auf dem Knote, wo das Minimum von *xerror* den relativen Fehler *rel error* plus den Standardfehler *xstd* nicht überschreitet. Die Erweiterung des Baumes hört hier deshalb bei $CP = 0.013$ auf.

	CP	nsplit	rel error	xerror	xstd
1	0.327	0	1.000	1.000	0.004
2	0.019	1	0.673	0.673	0.003
3	0.018	3	0.635	0.634	0.003
4	0.013	4	0.616	0.617	0.003
5	0.010	7	0.579	0.617	0.003

Tabelle 6.2: Komplexität Parameter

Stutzen des Baumes

Um den Ausgleich zwischen Genauigkeit und Komplexität zu finden, sollte der Regressionsbaum gestutzt (pruned) werden. Die grafische Darstellung der Kreuzvalidierungsschätzung in Abbildung 4.15 links enthält eine horizontale Linie, die dem Wert der 1 SE-Regel entspricht. Wie in Tabelle 4.3 gezeigt wird, an dem Knoten mit Komplexitätsparameter von 0.013 ist das Minimum erreicht (unter der horizontale Linie), ab diesem Knote sollte den Baum gestutzt werden. Die rechte Grafik zeigt die relativen Fehler gegen die Komplexitätsparameter für unterschiedliche Baumgrößen. Den gestutzten Regressionsbaum stellen wir in Abbildung 4.17 grafisch dar. Der Baum hat 4 Teilungen und die Ergebnisse erreichen 5 Blätter.

Interpretation

In Abbildung 4.17 sehen wir den Entscheidungsbaum nach dem Stutzen mit logarithmierte Bewertungsanzahl als die Zielgröße. Wir haben den Entscheidungsbaum in Abbildung 4.17 gewählt, weil er sich besser zu unseren Daten anpasst. Gelesen wird erst oben und

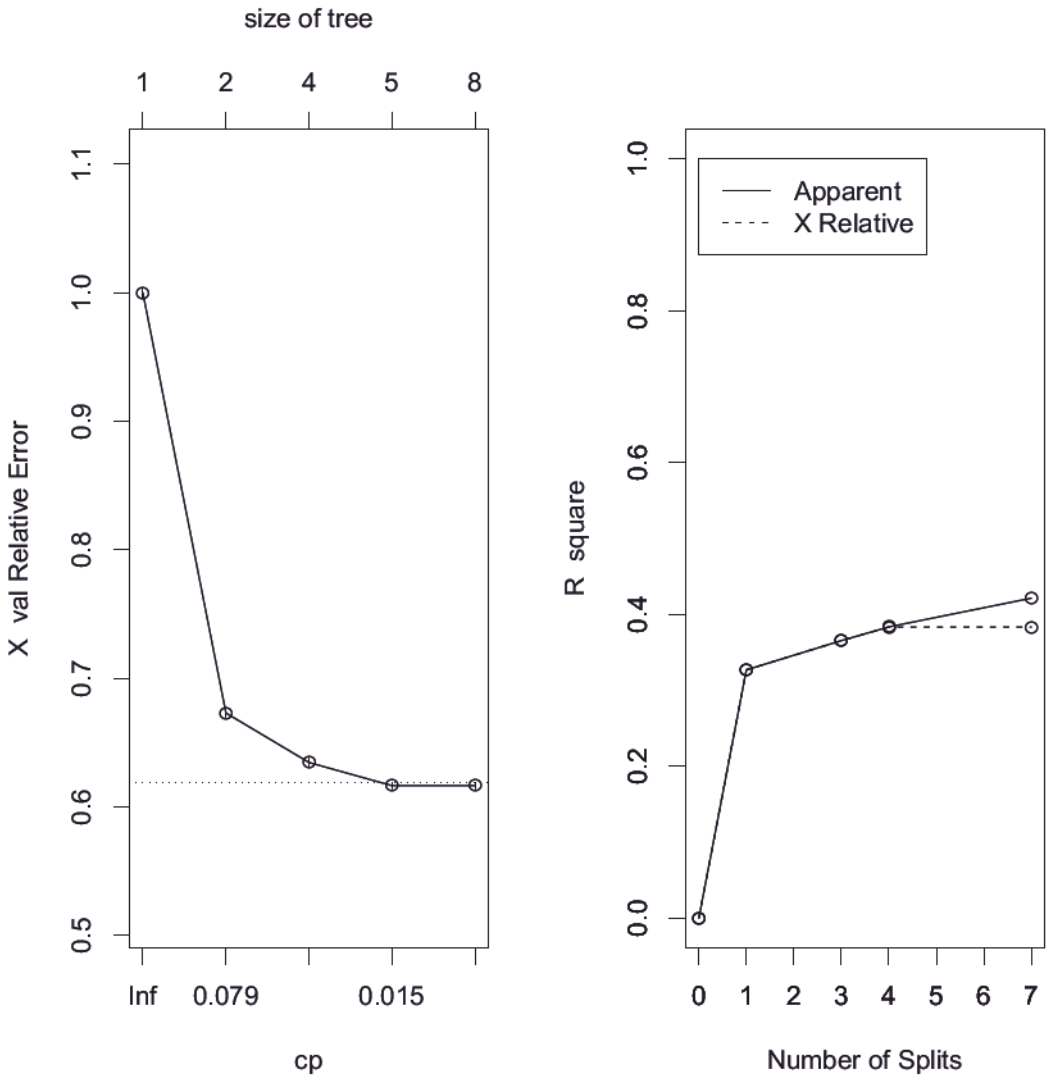


Abbildung 6.6: Trainieren des CART Modells

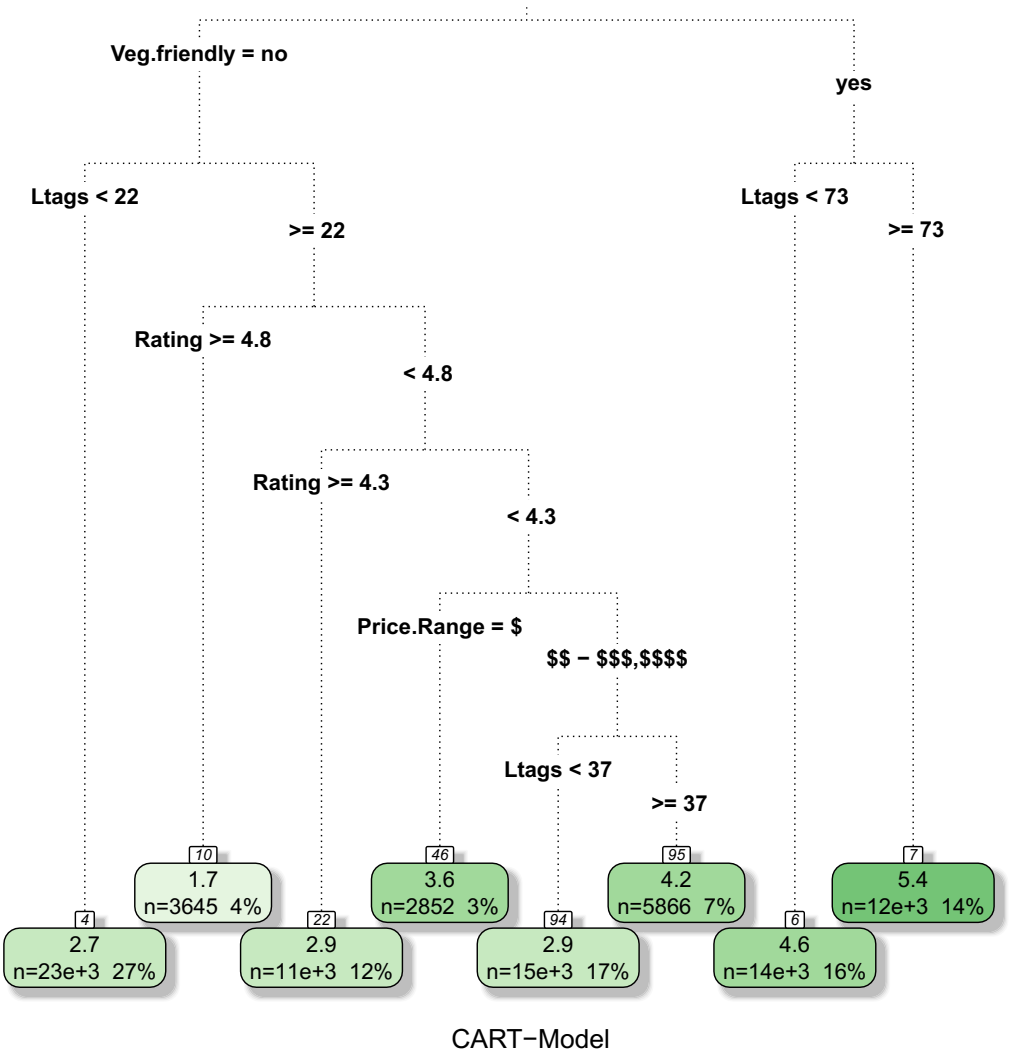


Abbildung 6.7: nicht-gestutzter Entscheidungsbaum

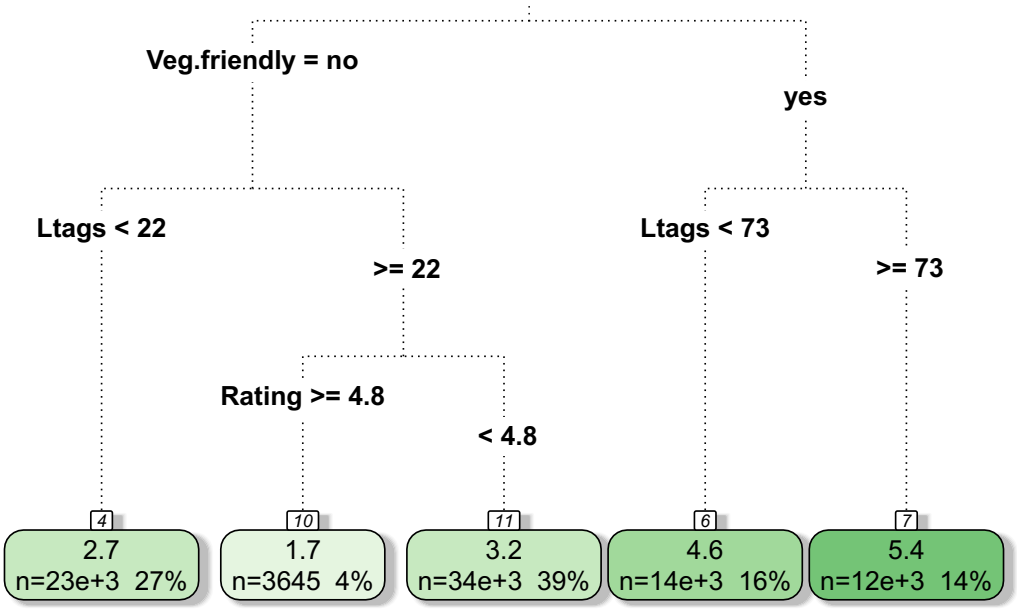


Abbildung 6.8: gestutzter Entscheidungsbaum

dann von links nach rechts. Im Folgenden stellen wir den gestutzten Baum im Zusammenhang mit dem Ergebnis der „summary“ - Funktion (siehe A.1 CART Output) dar.

Als wichtige Variable für die log-Anzahl der Bewertungen eines Restaurants dienen Veg. friendly (Vegetarier Freundlichkeit), Ltags (Beschreibungslänge) und Rating (Bewertungsniveau). Bei der Entwicklung des Baums werden 13871 NA-Werte ignoriert. In dem Wurzelknoten prüfen wir, ob das Restaurant Vegetarier freundlich ist. Bei Restaurants mit „Veg. friendly = no“ wird der linke Ast gewählt. Diese Knoten umfassen 70% der Trainingsdaten, die restlichen 30% Vegetarier freundliche Restaurants den rechten Ast. Auf der nächsten Stufe wird geprüft, ob „Rating“ ≥ 4.8 ist. Falls das Ergebnis positiv ist, wird der linken Ast gewählt. Unter dem rechten Knoten „Veg. friendly = yes“ werden zwei Äste geteilt: „Ltags < 73“ links und „Ltags ≥ 73“ recht. Die Blätter des Baumendes betrachtend können wir zusammenfassen, dass für nicht Vegetarier Restaurant die Wahrscheinlichkeiten von 27% für die log-Bewertungsanzahl von 2.7, 4% für 1.7 und 39% für 3.2 liefern würden, je nachdem ob sie weniger als 22 Zeichen über die Kochart beschreiben, oder mehr als 22 Zeichen, aberainer schlittgen r auf einer Bewertungsniveau von unter oder über 4.8 sind. Die Wahrscheinlichkeiten für nicht Vegetarier-freundliche Restaurants mit mehr als 73 Zeichen in der Beschreibung betragen 16% für 4.6 und 14% für 5.4 bei den Restaurants mit mehr als 73 Zeichen.

Es ergibt sich also ein deutlicher Unterschied zwischen Bewertungsanzahl der Restaurants mit und ohne „Vegetarian friendly“ in ihrer Beschreibung. Überraschend finden wir, dass die nicht Vegetarier-freundliche Restaurants trotz höchstem Bewertungsniveau $\geq$ 4.8 die niedrigste Bewertungsanzahl haben. Es ist irrelevant, auf welchem Preis- oder Bewertungsniveau sie sind, die Vegetarier-freundliche Restaurants haben allgemein höhere Anzahl der Bewertungen.

7. Fazit

Wir analysieren in dieser Arbeit Restaurant-Daten mit Hilfe statistischer Methoden. Um unseren Datensatz grafisch und empirisch zu analysieren nutzen wir die Programmiersprache R. Wir erklären den Zusammenhang zwischen verschiedenen Variablen indem wir ein multiples lineares Regressionsmodell sowie ein CART Regressionsmodell erstellen. Auf Grundlage der Datenanalyse zeigen wir, dass es die Länge der Selbstbeschreibung sowie die Selbstdarstellung als Vegetarier-freundlich sind, die einen deutlichen Unterschied in der Beliebtheit eines Restaurants machen.

Wir beginnen unsere Analyse damit, die wichtigsten quantifizierbaren Variablen für unsere Stichprobe auszuwählen. Wir betrachten und interpretieren fehlende Werte. In einer deskriptiven Analyse erfassen wir die statistischen Eigenschaften des Datensatzes und stellen ihn grafisch dar.

Anschließend betrachten wir die Korrelationen zwischen verschiedenen Variablen um einen Eindruck über lineare Zusammenhänge zu gewinnen. Wir arbeiten hierbei mit dem Spearman'schen Korrelationskoeffizienten. Wir sind überrascht festzustellen, dass zwischen der Anzahl der Bewertungen und dem Bewertungsniveau fast keine Korrelation entsteht. Dies erklären wir damit, dass das Bewertungsniveau sehr stabil in einem engen Intervall um 4.5 schwankt, völlig unabhängig von sonstigen beobachtbaren Variablen. Daher wählen wir uns für den Rest der Analyse die Anzahl der Bewertungen als Qualitätskriterium für ein Restaurant an Stelle des Bewertungsniveaus.

Entsprechend wählen wir für die multiple lineare Regression die logarithmische Anzahl der Bewertungen als abhängige Variable. Die Logarithmierung bewirkt eine stärkere Anpassung an eine Normalverteilung als die untransformierte Variable. Wir stellen eine starke Korrelation zwischen der Beschreibungslänge und der logarithmierten Anzahl der Bewertungen fest. Auch gibt es einen signifikanten Unterschied in der Bewertungsanzahlen zwischen Restaurants, die Vegetarierfreundlichkeit in ihrer Selbstbeschreibung erwähnen und solchen die hierüber keine Angabe machen.

Als nächstes untersuchen wir den Einfluss, den der Standort eines Restaurants ausübt. Hierzu fassen wir die Restaurants in einer Stadt zusammen und führen eine Clusteranalyse über Städte durch. Hieraus werden die Unterschiede und Ähnlichkeiten in Preisniveau und Anzahl der Bewertungen deutlich. Hierin erstellen wir ein hierarchisches Clustering sowie zwei k-means Clusterings. Für das hierarchische Clustering nutzen wir die Gower Distanz als Abstandsmaß zwischen Städten. Die Gower Distanz eignet sich besonders für gemischte Daten, die sowohl numerische als auch kategoriale Variablen enthalten. Wir erstellen das hierarchische Clustering auf Grundlage der Distanzmatrix mit Hilfe des Complete-Linkage Verfahrens. Auch mit K-Means Cluster Analyse werden die Unterschiede zwischen den Gruppen intuitiv darstellt.

Weiterhin stellen wir ein multiples lineares Regressionsmodell auf. Hierin ist die logarithmische Anzahl der Bewertungen die abhhängige Variable. Die unabhängigen Variablen Preiskategorie, Bewertungslänge, und Vegetarierfreundlichkeit haben allesamt signifikante positive Koeffizienten. Nur die unabhängige Variablen Bewertungsniveau hat einen signifikant negativen Koeffizienten. Das Bestimmtheitsmaß R^2 unseres multilinearen Modells ist mit 36% schlecht. Das kann entweder daran liegen, dass nicht alle

relevanten Variablen auch Teil des Datensatzes sind oder daran, dass der Zusammenhang zwischen den Variablen nicht linear ist.

Abschließend erstellen wir mit dem CART-Algorithmus einen Entscheidungsbaum. Wir nutzen den Entscheidungsbaum um unbekannte Variablen vorherzusagen. Besonders sinnvoll ist nach unserer Einschätzung für die Vorhersage einer zu erwartenden Anzahl von Bewertungen, dass man die Selbstbeschreibung und die Preiskategorie eines Restaurants kennt. Mit dem Komplexitätsparameter steuern wir die Anpassungsgüte des Modells. Hierbei unterscheiden wir die Anpassung an den Trainingsdatensatz und die Anpassung an einen hypothetischen neuen Datensatz. Wir führen eine n-fache Kreuzvalidierung durch um den optimalen Komplexitätsparameter zu finden und stutzen den Entscheidungsbaum zu sodass er die gewünschte Komplexität besitzt. Die Struktur des Entscheidungsbaums erlaubt hierin Rückschlüsse darauf, welche Variablen die größte Information tragen. Die wichtigste Information ist hierbei ob Vegetarier-Freundlichkeit erwähnt wird oder nicht. Die zweit-wichtigste Information ist die Länge der Beschreibung. Für Restaurants, die über ihre Vegetarier-Freundlichkeit keine Angabe machen und eine Selbstbeschreibung haben, die länger oder gleich 22 Zeichen sind ist das Bewertungsniveau eine Diskriminante. Das Preisniveau spielt in dem konstruierten Entscheidungsbaum keine Rolle.

Literatur

[Alp10] Ethem Alpaydin. Introduction to machine learning. [sl], 2010.

[BEW15] Klaus Backhaus, Bernd Erichson, and Rolf Weiber. *Fortgeschrittene multivariate Analysemethoden: eine anwendungsorientierte Einführung*. Springer-Verlag, 2015.

[FKPT11] L Fahrmeir, R Künstler, I Pigeot, and G Tutz. Statistik der weg zur datenanalyse (7. aufl., korr. nachdr.). *Berlin [ua]: Springer*, 2011.

[HK17] A. Handl and T. Kuhlenkasper. *Multivariate Analysemethoden: Theorie und Praxis mit R*. Statistik und ihre Anwendungen. Springer Berlin Heidelberg, 2017.

[KS16] Lasse Kliemann and Peter Sanders. *Algorithm Engineering: Selected Results and Surveys*, volume 9220. Springer, 2016.

[Sch13] Rainer Schlittgen. *Regressionsanalysen mit R*. Walter de Gruyter, 2013.

[TA19] TM Therneau and EJ Atkinson. An introduction to recursive partitioning using the rpart routines; 2015. *URL: https://cran. r-project. org/web/packages/rpart/vignettes/longintro. pdf (date of access: 21.11. 2018)*, 2019.

[Wun14] Johannes Wunder. *Analyse des Verhaltens verschiedener Clusterverfahren nach Imputation fehlender Daten*. PhD thesis, 2014.

[ZMH07] Achim Zeileis, David Meyer, and Kurt Hornik. Residual-based shadings for visualizing (conditional) independence. *Journal of Computational and Graphical Statistics*, 16(3):507–525, 2007.

A. R-Outputs

A.1. Beschreibungslänge

```
##table(D$Ltags)

   7    8    9   10   11   12   13   14   15   16   17   18   19 20
 630 1903 2779 3814 8533 2310  797 1406  129  288 1915  562  941 2214
  21   22   23   24   25   26   27   28   29   30   31   32   33 34
2180 2471 2077  933 1495 1809  392 1985  984 1081  910 1010 1122 1293
  35   36   37   38   39   40   41   42   43   44   45   46   47 48
 897  685 1062  745  871 1660  619 1044 1521  642 1044  790  574 866
  49   50   51   52   53   54   55   56   57   58   59   60   61 62
 840  598 1241  701  771  738  873  542  731  488  679  942  689 696
  63   64   65   66   67   68   69   70   71   72   73   74   75 76
1539  516  635  622  368  593  515  476  559  986  420  781  473 405
  77   78   79   80   81   82   83   84   85   86   87   88   89 90
 504  393  316  642  329  555  612  441  511  768  270  337  519 208
  91   92   93   94   95   96   97   98   99  100  101  102  103 104
 451  279  352  306  354  253  228  221  133  338  239  252  674 172
 105  106  107  108  109  110  111  112  113  114  115  116  117 118
 132  202  128  130  150  119  198  340  121  175  106   58   68 77
 119  120  121  122  123  124  125  126  127  128  129  130  131 132
 150   94   78   70   95   52   63   27   33   55   23   68   35 35
 133  134  135  136  137  138  139  140  141  142  143  144  145 146
  17   21   12   11   15   13   18   12   12    8    5    5 21    4
 147  149  150  151  152  153  154  155  156  157  158  159  160 161
   9    5    4    2    2    4    5    4    6    2    1    1 1    1
 162  164  166  169  170  171  172  180  203  207  263
   2    2    1    1    1    3    1    2    1    1    1
```

A.2. LM Output

```
##summary(m)

Call:
lm(formula = log(Number.of.Reviews) ~ Rating + Pcat + Ltags +
    Veg, data = D)

Residuals:
    Min      1Q  Median      3Q     Max
-5.3801 -0.7237 -0.0036  0.7434  5.9554

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.917879   0.034818  112.52   <2e-16 ***
Rating      -0.421761   0.007766  -54.31   <2e-16 ***
Pcat         0.455524   0.008374   54.40   <2e-16 ***
Ltags        0.019360   0.000209   92.65   <2e-16 ***
Veg          0.554656   0.012389   44.77   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.14 on 74234 degrees of freedom
  (51288 observations deleted due to missingness)
Multiple R-squared:  0.3629,    Adjusted R-squared:  0.3628
F-statistic: 1.057e+04 on 4 and 74234 DF,  p-value: < 2.2e-16
```

A.3. CART Output

```
## rsq.rpart(cm)

Regression tree:
rpart(formula = log(Number.of.Reviews) ~ ., data = train, control = rpart.control(minsplit = 20000,
    minbucket = 1000, CP = 0.001))
r
Variables actually used in tree construction:
[1] Ltags        Price.Range  Rating       Veg.friendly

Root node error: 236569/86550 = 2.7333
:
n=86550 (13871 observations deleted due to missingness)

        CP nsplit rel error  xerror      xstd
1 0.326954      0   1.00000 1.00002 0.0038688
2 0.019264      1   0.67305 0.67307 0.0030623
3 0.018069      3   0.63452 0.63457 0.0028906
4 0.012633      4   0.61645 0.61651 0.0028499
5 0.010000      7   0.57855 0.61651 0.0028499

## summary(cm)
Call:
rpart(formula = log(Number.of.Reviews) ~ ., data = train, control = rpart.control(minsplit = 20000,
    minbucket = 1000, CP = 0.001))
  n=86550 (13871 observations deleted due to missingness)

          CP nsplit rel error    xerror        xstd
1 0.32695375      0 1.0000000 1.0000170 0.003868821
2 0.01926441      1 0.6730462 0.6730724 0.003062339
3 0.01806924      3 0.6345174 0.6345677 0.002890571
4 0.01263327      4 0.6164482 0.6165061 0.002849863
5 0.01000000      7 0.5785484 0.6165061 0.002849863

Variable importance
Veg.friendly        Ltags       Rating  Price.Range
          76           14            9            1

Node number 1: 86550 observations,    complexity param=0.3269538
  mean=3.513529, MSE=2.733319
  left son=2 (60834 obs) right son=3 (25716 obs)
  Primary splits:
      Veg.friendly splits as  LR,       improve=0.32695380, (0 missing)
      Ltags        < 42.5 to the left,  improve=0.19075000, (18136 missing)
      Rating       < 4.75 to the right, improve=0.06210304, (0 missing)
      Price.Range  splits as  LRR,      improve=0.03288550, (27142 missing)

Node number 2: 60834 observations,    complexity param=0.01926441
  mean=2.898895, MSE=2.009234
  left son=4 (23059 obs) right son=5 (37775 obs)
  Primary splits:
      Ltags       < 21.5 to the left,  improve=0.07546125, (18136 missing)
      Rating      < 4.75 to the right, improve=0.06401102, (0 missing)
      Price.Range splits as  LRR,      improve=0.02302297, (26894 missing)
  Surrogate splits:
      Rating < 3.25 to the left,  agree=0.548, adj=0.034, (18136 split)

Node number 3: 25716 observations,    complexity param=0.01806924
  mean=4.967513, MSE=1.43848
  left son=6 (13809 obs) right son=7 (11907 obs)
  Primary splits:
```

```
      Ltags       < 72.5 to the left,  improve=0.11555530, (0 missing)
      Rating      < 4.75 to the right, improve=0.05478327, (0 missing)
      Price.Range splits as  LRR,      improve=0.04961147, (248 missing)
  Surrogate splits:
      Rating      < 4.25 to the left,  agree=0.568, adj=0.068, (0 split)
      Price.Range splits as  LLR,      agree=0.547, adj=0.022, (0 split)

Node number 4: 23059 observations
  mean=2.65395, MSE=1.507124

Node number 5: 37775 observations,    complexity param=0.01926441
  mean=3.048416, MSE=2.256756
  left son=10 (3645 obs) right son=11 (34130 obs)
  Primary splits:
      Rating      < 4.75 to the right, improve=0.08078335, (0 missing)
      Price.Range splits as  LRR,      improve=0.02170412, (17442 missing)
      Ltags       < 36.5 to the left,  improve=0.01104320, (15062 missing)

Node number 6: 13809 observations
  mean=4.588925, MSE=1.243754

Node number 7: 11907 observations
  mean=5.406576, MSE=1.30531

Node number 10: 3645 observations
  mean=1.741876, MSE=0.803296

Node number 11: 34130 observations,    complexity param=0.01263327
  mean=3.187951, MSE=2.210204
  left son=22 (10575 obs) right son=23 (23555 obs)
  Primary splits:
      Rating      < 4.25 to the right, improve=0.02147682, (0 missing)
      Price.Range splits as  LRR,      improve=0.01828854, (15030 missing)
      Ltags       < 36.5 to the left,  improve=0.01332388, (12982 missing)

Node number 22: 10575 observations
  mean=2.862787, MSE=2.149387

Node number 23: 23555 observations,    complexity param=0.01263327
  mean=3.333934, MSE=2.168728
  left son=46 (2852 obs) right son=47 (20703 obs)
  Primary splits:
      Price.Range splits as  LRR,      improve=0.015433230, (9577 missing)
      Ltags       < 34.5 to the left,  improve=0.013334300, (8164 missing)
      Rating      < 3.25 to the right, improve=0.006409523, (0 missing)

Node number 46: 2852 observations
  mean=3.641505, MSE=1.468704

Node number 47: 20703 observations,    complexity param=0.01263327
  mean=3.291563, MSE=2.250335
  left son=94 (14837 obs) right son=95 (5866 obs)
  Primary splits:
      Ltags  < 36.5 to the left,  improve=0.013127150, (8164 missing)
      Rating < 3.25 to the right, improve=0.008957147, (0 missing)
  Surrogate splits:
      Rating < 1.25 to the right, agree=0.533, adj=0.002, (8164 split)

Node number 94: 14837 observations
  mean=2.92493, MSE=1.964612

Node number 95: 5866 observations
```

```
  mean=4.218898, MSE=1.77308

## cor(predict(cm, newdata=na.omit(test)), log((na.omit(test))$Number.of.Reviews))^2
[1] 0.3194276
## cor(predict(pruned_cm, newdata=na.omit(test)), log((na.omit(test))$Number.of.Reviews))^2
[1] 0.3181712
```

Erklärung zur Urheberschaft

Hiermit erkläre ich, Xiaoji Du, dass ich die vorliegende Arbeit noch nicht für andere Prüfungen eingereicht habe. Ich habe die Arbeit selbständig verfasst. Sämtliche Quellen einschließlich Internetquellen, die ich unverändert oder abgewandelt wiedergegeben habe, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, habe ich als solche kenntlich gemacht. Ich bin mir darüber bewusst, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.